Cross-Layer Designs in Coded Wireless Fading Networks with Multicast

Ketan Rajawat, Nikolaos Gatsis, and Georgios B. Giannakis, *Fellow, IEEE*

Abstract—A cross-layer design along with an optimal resource allocation framework is formulated for wireless fading networks, where the nodes are allowed to perform network coding. The aim is to jointly optimize end-to-end transport layer rates, network code design variables, broadcast link flows, link capacities, average power consumption, and short-term power allocation policies. As in the routing paradigm where nodes simply forward packets, the cross-layer optimization problem with network coding is non-convex in general. It is proved however, that with network coding, dual decomposition for multicast is optimal so long as the fading at each wireless link is a continuous random variable. This lends itself to provably convergent subgradient algorithms, which not only admit a layered-architecture interpretation but also optimally integrate network coding in the protocol stack. The dual algorithm is also paired with a scheme that yields near-optimal network design variables, namely multicast end-to-end rates, network code design quantities, flows over the broadcast links, link capacities, and average power consumption. Finally, an asynchronous subgradient method is developed, whereby the dual updates at the physical layer can be affordably performed with a certain delay with respect to the resource allocation tasks in upper layers. This attractive feature is motivated by the complexity of the physical layer subproblem, and is an adaptation of the subgradient method suitable for network control.

Index Terms—Network coding, cross-layer designs, optimization methods, asynchronous subgradient methods, multihop.

I. INTRODUCTION

Traditional networks have always assumed nodes capable of only forwarding or replicating packets. For many types of networks however, this constraint is not inherently needed since the nodes can invariably perform encoding functions. Interestingly, even simple linear mixing operations can be powerful enough to enhance the network throughput, minimize delay, and decrease the overall power consumption [1], [2]. For the special case of single-source multicast, which does not even admit a polynomial-time solution within the routing framework [3], linear network coding achieves the full network capacity [4]. In fact, the network flow description of multicast with random network coding adheres to only linear inequality constraints reminiscent of the corresponding description in unicast routing [5].

This encourages the use of network coding to extend several popular results in unicast routing framework to multicast without appreciable increase in complexity. Of particular interest is the resource allocation and cross-layer optimization task in wireless networks [6], [7]. The objective here is to maximize

a network utility function subject to flow, rate, capacity and power constraints. This popular approach not only offers the flexibility of capturing diverse performance objectives, but also admits a layering interpretation, arising from different decompositions of the optimization problem [8].

This paper deals with cross-layer optimization of wireless *multicast* networks that use network coding and operate over *fading* links. The aim is to maximize a total network utility objective, and entails finding end-to-end rates, network code design variables, broadcast link flows, link capacities, average power consumption, and instantaneous power allocations.

Network utility maximization was first brought into coded networks in [5], where the aim was to minimize a generic cost function subject only to flow and rate constraints. The optimal flow and rate variables may then be converted to a practical random network coding implementation using methods from [9] and [10]. Subsequent works extended this framework to include power, capacity, and scheduling constraints [11]–[14]. The interaction of network coding with the network and transport layers has also been explored in [15]–[19]; in these works, networks with fixed link capacities are studied, and different decomposition techniques result in different types of layered architectures.

There are however caveats associated with the utility maximization problem in wireless networks. First, the power control and scheduling subproblems are usually non-convex. This implies that the dual decomposition of the overall problem, though insightful, is not necessarily optimal and does not directly result in a feasible primal solution. Second, for continuous fading channels, determining the power control policy is an infinite dimensional problem. Existing approaches in network coding consider either deterministic channels [11], [14], or, links with a finite number of fading states [12], [20], [21].

On the other hand, a recent result in unicast routing shows that albeit the non-convexity, the overall utility optimization problem has no duality gap for wireless networks with continuous fading channels [22]. As this is indeed the case in all real-life fading environments, the result promises the optimality of layer separation. In particular, it renders a dual subgradient descent algorithm for network design optimal [23].

The present paper begins with a formulation that jointly optimizes end-to-end rates, virtual flows, broadcast link flows, link capacities, average power consumption, and instantaneous power allocations in wireless fading multicast networks that use intra-session network coding (Section II). The first contribution of this paper is to introduce a *realistic physical layer model* formulation accounting for the capacity of broadcast links. The cross-layer problem is generally non-convex, yet it is shown to have zero duality gap (Section III-A). This result considerably broadens [22] to *coded multicast* networks with

The authors are with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455, USA. Tel/fax: (612)624-9510/625-2002, emails: {ketan,gatsisn,georgios}@umn.edu

broadcast links. The zero duality gap is then leveraged in order to develop a subgradient descent algorithm that minimizes the dual function (Sections III-B, III-C). The algorithm admits a natural layering interpretation, allowing optimal integration of network coding into the protocol stack.

In Section IV, the subgradient algorithm is modified so that the component of the subgradient that results from the physical layer power allocation may be delayed with respect to operations in other layers. This provably convergent asynchronous subgradient method and its *online* implementation constitute the second major contribution. Unlike the algorithm in [23], which is used for offline network optimization, the algorithm developed here is suitable for online network control. Convergence of asynchronous subgradient methods for dual minimization is known under diminishing stepsize [24]; the present paper proves results for constant stepsize. Near-optimal primal variables are also recovered by forming running averages of the primal iterates. This technique has also been used in synchronous subgradient methods for convex optimization; see e.g., [25] and references therein. Here, ergodic convergence results are established for the asynchronous scheme and the non-convex problem at hand. Finally, numerical results are presented in Section V, and Section VI concludes the paper.

II. PROBLEM FORMULATION

Consider a wireless network consisting of a set of terminals (nodes) denoted by $\mathcal{N}$. The broadcast property of the wireless interface is modeled by using the concept of hyperarcs. A *hyperarc* is a pair (i, J) that represents a broadcast link from a node i to a chosen set of nodes $J \subset \mathcal{N}$. The entire network can therefore be represented as a hypergraph $\mathcal{H} = (\mathcal{N}, \mathcal{A})$, where $\mathcal{A}$ is the set of hyperarcs. The complexity of the model is determined by the choice of the set $\mathcal{A}$. Let the neighbor-set $N(i)$ denote the set of nodes that node i reaches. An exhaustive model might include all possible $2^{|N(i)|} - 1$ hyperarcs from node i. On the other hand, a simpler model might include only a smaller number of hyperarcs per node. A point-to-point model is also a special case when node i has $|N(i)|$ hyperarcs each containing just one receiver.

The present work considers a physical layer whereby the channels undergo random multipath fading. This model allows for opportunistically best schedules per channel realization. This is different from the link-level network models in [5], [12], [13], [21], where the hyperarcs are modeled as erasure channels. The next subsection discusses the physical layer model in detail.

A. Physical Layer

In the current setting, terminals are assumed to have a set of tones $\mathcal{F}$ available for transmission. Let h_{ij}^f denote the power gain between nodes i and j over a tone $f \in \mathcal{F}$, assumed random, capturing fading effects. Let $\mathbf{h}$ represent the vector formed by stacking all the channel gains. The network operates in a time slotted fashion; the channel $\mathbf{h}$ remains constant for the duration of a slot, but is allowed to change from slot to slot. A slowly fading channel is assumed so that a large number of packets may be transmitted per time slot. The fading process is modeled to be stationary and ergodic.

Since the channel changes randomly per time slot, the optimization variables at the physical layer are the channel realization-specific power allocations $p_{iJ}^f(\mathbf{h})$ for all hyperarcs $(i, J) \in \mathcal{A}$, and tones $f \in \mathcal{F}$. For convenience, these power allocations are stacked in a vector $\mathbf{p}(\mathbf{h})$. Instantaneous power allocations may adhere to several scheduling and mask constraints, and these will be generically denoted by a bounded set Π such that $\mathbf{p}(\mathbf{h}) \in \Pi$. The long-term average power consumption by a node i is given by

$$p_i = \mathbb{E}\left[\sum_f \sum_{J:(i,J)\in\mathcal{A}} p_{iJ}^f(\mathbf{h})\right] \qquad (1)$$

where $\mathbb{E}[.]$ denotes expectation over the stationary channel distribution.

For slow fading channels, the information-theoretic capacity of a hyperarc (i, J) is defined as the maximum rate at which *all* nodes in J receive data from i with vanishing probability of error in a given time slot. This capacity depends on the instantaneous power allocations $\mathbf{p}(\mathbf{h})$ and channels $\mathbf{h}$. A generic bounded function $C_{iJ}^f(\mathbf{p}(\mathbf{h}), \mathbf{h})$ will be used to describe this mapping. Next we give two examples of the functional forms of $C_{iJ}^f(\cdot)$ and Π.

Example 1. Conflict graph model: The power allocations p_{iJ}^f adhere to the spectral mask constraints

$$0 \le p_{iJ}^f \le p_{\max}^f. \qquad (2)$$

However, only conflict-free hyperarc are allowed to be scheduled for a given $\mathbf{h}$. Specifically, power may be allocated to hyperarcs (i_1, J_1) and (i_2, J_2) if and only if [13]

i) $i_1 \neq i_2$;
ii) $i_1 \notin J_2$ and $i_2 \notin J_1$ (half-duplex operation); and
iii-a) $J_1 \cap J_2 = \emptyset$ (primary interference), or additionally,
iii-b) $J_1 \cap N(i_2) = J_2 \cap N(i_1) = \emptyset$ (secondary interference).
The set Π therefore consists of all possible power allocations that satisfy the previous properties.

Due to hyperarc scheduling, all transmissions in the network are interference free. The signal-to-noise ratio (SNR) at a node $j \in J$ is given by

$$\Gamma_{iJj}^f(\mathbf{p}(\mathbf{h}), \mathbf{h}) = p_{iJ}^f(\mathbf{h}) h_{ij}^f / N_j \qquad (3)$$

where N_j is the noise power at j. In a broadcast setting, the maximum rate of information transfer from i to *each* node in J is

$$C_{iJ}^f(\mathbf{p}(\mathbf{h}), \mathbf{h}) = \min_{j \in J} \log(1 + \Gamma_{iJj}^f(\mathbf{p}(\mathbf{h}), \mathbf{h})). \qquad (4)$$

A similar expression can be written for the special case of point-to-point links by substituting hyperarcs (i, J) by arcs (i, j) in the expression for $\Gamma_{iJj}^f(\mathbf{p}(\mathbf{h}), \mathbf{h})$.

For slow-fading channels, Gaussian codebooks with sufficiently large block lengths achieve this capacity in every time slot. More realistically, an SNR penalty term ρ can be included to account for finite-length practical codes and adaptive modulation schemes, so that

$$C_{iJ}^f(\mathbf{p}(\mathbf{h}), \mathbf{h}) = \min_{j \in J} \log\left(1 + \Gamma_{iJj}^f(\mathbf{p}(\mathbf{h}), \mathbf{h})/\rho\right). \qquad (5)$$

The penalty term is in general a function of the target bit error rate.

Example 2. Signal-to-interference-plus-noise-ratio (SINR) model: Here, the constraint set Π is simply a box set $\mathcal{B}_{\mathbf{p}}$,

$$\Pi = \mathcal{B}_{\mathbf{p}} := \{p_{iJ}^f | 0 \leq p_{iJ}^f \leq p_{\max}^f \ \forall \ (i, J) \in \mathcal{A} \text{ and } f \in \mathcal{F} \}. \tag{6}$$

The set $\mathcal{B}_{\mathbf{p}}$ could also include (instantaneous) sum-power constraints per node. The capacity is expressed as in (4) or (5), but now the SNR is replaced by the SINR, given by

$$\Gamma_{iJj}^f(\mathbf{p}(\mathbf{h}), \mathbf{h}) = p_{iJ}^f(\mathbf{h})h_{ij}^f \Big/ \left(N_j + I_{ij,f}^{\text{int}} + I_{j,f}^{\text{self}} + I_{iJj,f}^{\text{broad}} \right). \tag{7}$$

The denominator consists of the following terms:

- Interference from other nodes' transmissions to node j

$$I_{ij,f}^{\text{int}} = \sum_{\substack{(k,M)\in\mathcal{A}:j\in M, \\ k\neq j, k\neq i}} p_{kM}^f(\mathbf{h})h_{kj}^f. \tag{8a}$$

- "Self-interference" due to transmissions of node j

$$I_{j,f}^{\text{self}} = h_{jj} \sum_{M:(j,M)\in\mathcal{A}} p_{jM}^f(\mathbf{h}). \tag{8b}$$

  This term is introduced to encourage half-duplex operation by setting h_{jj} to a large value.

- "Broadcast-interference" from transmissions of node i to other hyperarcs

$$I_{iJj,f}^{\text{broad}} = \beta h_{ij}^f \sum_{\substack{M:(i,M)\in\mathcal{A} \\ M\neq J}} p_{iM}^f(\mathbf{h}). \tag{8c}$$

  This term is introduced to force node i to transmit at most over a single hyperarc, by setting β to a large value.

The previous definitions ignore interference from non-neighboring nodes. However, they can be readily extended to include more general interference models.

The link layer capacity is defined as the long-term average of the total instantaneous capacity, namely,

$$c_{iJ} := \mathbb{E}\left[\sum_f C_{iJ}^f(\mathbf{p}(\mathbf{h}), \mathbf{h}) \right]. \tag{9}$$

This is also called ergodic capacity and represents the maximum average data rate available to the link layer.

B. Link Layer and Above

The network supports multiple multicast sessions indexed by m, namely $\mathcal{S}_m := (s^m, T^m, a^m)$, each associated with a source node s^m, sink nodes $T^m \subset \mathcal{N}$, and an average flow rate a^m from s^m to each $t \in T^m$. The value a^m is the average rate at which the network layer of source terminal s^m admits packets from the transport layer. Traffic is considered elastic, so that the packets do not have any short-term delay constraints.

Network coding is a generalization of routing since the nodes are allowed to code packets together rather than simply forward them. This paper considers intra-session network coding, where only the traffic belonging to the same multicast session is allowed to mix. Although better than routing in general, this approach is still suboptimal in terms of achieving the network capacity. However, general (inter-session) network coding is difficult to characterize or implement since neither the capacity region nor efficient network code designs are known [1, Part II]. On the other hand, a simple linear coding strategy achieves the full capacity region of intra-session network coding [4].

The network layer consists of endogenous flows of coded packets over hyperarcs. Recall that the maximum average rate of transmission over a single hyperarc cannot exceed c_{iJ}. Let the coded packet-rate of a multicast session m over hyperarc (i, J) be z_{iJ}^m (also referred to as the subgraph or broadcast link flow). The link capacity constraints thus translate to

$$\sum_m z_{iJ}^m \leq c_{iJ} \quad \forall (i, J) \in \mathcal{A}. \tag{10}$$

To describe the intra-session network coding capacity region, it is commonplace to use the concept of *virtual flow* between terminals i and j corresponding to each session m and sink $t \in T^m$ with average rate x_{ij}^{mt}. These virtual flows are defined only for neighboring pairs of nodes i.e., $(i, j) \in \mathcal{G} := \{(i,j)|(i, J) \in \mathcal{A}, j \in J\}$. The virtual flows satisfy the flow-conservation constraints, namely,

$$\sum_{j:(i,j)\in\mathcal{G}} x_{ij}^{mt} - \sum_{j:(j,i)\in\mathcal{G}} x_{ji}^{mt} = \sigma_i^m := \begin{cases} a^m & \text{if } i = s^m, \\ -a^m & \text{if } i = t, \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

for all m, $t \in T^m$, and $i \in \mathcal{N}$. Hereafter, the set of equations for $i = t$ will be omitted because they are implied by the remaining equations.

The broadcast flows z_{iJ}^m and the virtual flows x_{ij}^{mt} can be related using results from the lossy-hyperarc model of [5], [13]. Specifically, [13, eq. (9)] relates the virtual flows and subgraphs, using the fraction $b_{iJK} \in [0, 1]$ of packets injected into the hyperarc (i, J) that reach the set of nodes $K \subset N(i)$. Recall from Section II-A, that here the instantaneous capacity function $C_{iJ}^f(\cdot)$ is defined such that all packets injected into the hyperarc (i, J) are received by every node in J. Thus in our case, $b_{iJK} = 1$ whenever $K \cap J \neq \emptyset$ and consequently,

$$\sum_{j\in K} x_{ij}^{mt} \leq \sum_{\substack{J:(i,J)\in\mathcal{A} \\ J\cap K\neq\emptyset}} z_{iJ}^m \quad \forall K \subset N(i), i \in \mathcal{N}, m, t \in T^m. \tag{12}$$

Note the difference with [13] where at every time slot, packets are injected into a fixed set of hyperarcs at the same rate. The problem in [13] is therefore to find a schedule of hyperarcs that do not interfere (the non-conflicting hyperarcs). The same schedule is used at every time slot; however, only a random subset of nodes receive the injected packets in a given slot. Instead here, the hyperarc selection is part of the power allocation problem at the physical layer, and is done for every time slot. The transmission rate (or equivalently, the channel coding redundancy) is however appropriately adjusted so that all the nodes in the selected hyperarc receive the data.

In general, for any feasible solution to the set of equations (10)-(12), a network code exists that supports the corresponding exogenous rates a^m [5]. This is because for each multicast session m, the maximum flow between s^m and $t \in T^m$ is a^m, and is therefore achievable [4, Th. 1]. Given a feasible solution, various network coding schemes can be used to achieve the exogenous rates. Random network coding based

implementations such as those proposed in [9] and [10], are particularly attractive since they are fully distributed and require little overhead. These schemes also handle any residual errors or erasures that remain due to the physical layer.

The system model also allows for a set of "box constraints" that limit the long-term powers, transport layer rates, broadcast link flow rates, virtual flow rates as well as the maximum link capacities. Combined with the set Π, these constraints can be compactly expressed as

$$
\begin{aligned}
\mathcal{B} := \{ \mathbf{y}, \mathbf{p}(\mathbf{h}) | \ \mathbf{p}(\mathbf{h}) \in \Pi, \quad & 0 \le p_i \le p_i^{\max}, \\
a_{\min}^m \le a^m \le a_{\max}^m, \quad & 0 \le c_{iJ} \le c_{iJ}^{\max}, \\
0 \le z_{iJ}^m \le z_{iJ}^{\max}, \quad & 0 \le x_{ij}^{mt} \le x_{ij}^{\max} \}.
\end{aligned}
\tag{13}
$$

Here $\mathbf{y}$ is a super-vector formed by stacking all the average rate and power variables, that is, a^m, z_{iJ}^m, x_{ij}^{mt}, c_{iJ}, and p_i. Parameters with min/max subscripts or superscripts denote prescribed lower/upper bounds on the corresponding variables.

C. Optimal Resource Allocation

A common objective of the network optimization problem is maximization of the exogenous rates a^m and minimization of the power consumption p_i. Towards this end, consider increasing and concave utility functions $U_m(a^m)$ and convex cost functions $V_i(p_i)$ so that the overall objective function $f(\mathbf{y}) = \sum_m U_m(a^m) - \sum_i V(p_i)$ is concave. For example, the utility function can be the logarithm of session rates and the cost function can be the squared average power consumption. The network utility maximization problem can be written as

$$
\mathsf{P} = \max_{(\mathbf{y}, \mathbf{p}(\mathbf{h})) \in \mathcal{B}} \sum_m U_m(a^m) - \sum_i V_i(p_i)
\tag{14a}
$$

$$
\text{s. t. } \sigma_i^m \le \sum_{(i,j) \in \mathcal{G}} x_{ij}^{mt} - \sum_{(j,i) \in \mathcal{G}} x_{ji}^{mt} \quad \forall \ m, i \ne t, t \in T^m
\tag{14b}
$$

$$
\sum_{j \in K} x_{ij}^{mt} \le \sum_{\substack{J:(i,J) \in \mathcal{A} \\ J \cap K \ne \emptyset}} z_{iJ}^m \quad \forall \ K \subset N(i), m, t \in T^m
\tag{14c}
$$

$$
\sum_m z_{iJ}^m \le c_{iJ} \quad \forall \ (i,J) \in \mathcal{A}
\tag{14d}
$$

$$
c_{iJ} \le \mathbb{E}\left[\sum_f C_{iJ}^f(\mathbf{p}(\mathbf{h}), \mathbf{h}) \right] \quad \forall \ (i,J) \in \mathcal{A}
\tag{14e}
$$

$$
\mathbb{E}\left[\sum_f \sum_{J:(i,J) \in \mathcal{A}} p_{iJ}^f(\mathbf{h}) \right] \le p_i
\tag{14f}
$$

where $i \in \mathcal{N}$. Note that constraints (1), (9) and (11) have been relaxed without increasing the objective function. For instance, the relaxation of (11) is equivalent to allowing each node to send at a higher rate than received, which amounts to adding virtual sources at all nodes $i \ne t$. However, adding virtual sources does not result in an increase in the objective function because the utilities U_m depend only on the multicast rate a^m.

The solution of the optimization problem (14) gives the throughput a^m that is achievable using optimal virtual flow rates x_{ij}^{mt} and power allocation policies $\mathbf{p}(\mathbf{h})$. These virtual flow rates are used for network code design. When implementing coded networks in practice, the traffic is generated in

packets and stored at nodes in queues (and virtual queues for virtual flows) [10]. The constraints in (14) guarantee that all queues are stable.

Optimization problem (14) is non-convex in general, and thus difficult to solve. For example, in the conflict graph model, the constraint set Π is discrete and non-convex, while in the SINR-model, the capacity function $C_{iJ}^f(\mathbf{p}(\mathbf{h}), \mathbf{h})$ is a non-concave function of $\mathbf{p}(\mathbf{h})$; see e.g., [26], [6]. The next section analyzes the Lagrangian dual of (14).

III. OPTIMALITY OF LAYERING

This section shows that (14) has zero duality gap, and solves the dual problem via subgradient descent iterations. The purpose here is two-fold: i) to describe a layered architecture in which linear network coding is optimally integrated; and ii) to set the basis for a network implementation of the subgradient method, which will be developed in Section IV.

A. Duality Properties

Associate Lagrange multipliers ν_i^{mt}, η_{iK}^{mt}, ξ_{iJ}, λ_{iJ} and μ_i with the flow constraints (14b), the union of flow constraints (14c), the link rate constraints (14d), the capacity constraints (14e), and the power constraints (14f), respectively. Also, let $\boldsymbol{\zeta}$ be the vector formed by stacking these Lagrange multipliers in the aforementioned order. Similarly, if inequalities (14b)–(14f) are rewritten with zeros on the right-hand side, the vector $\mathbf{q}(\mathbf{y}, \mathbf{p}(\mathbf{h}))$ collects all the terms on the left-hand side of the constraints. The Lagrangian can therefore be written as

$$
\mathcal{L}(\mathbf{y}, \mathbf{p}(\mathbf{h}), \boldsymbol{\zeta}) = \sum_m U_m(a^m) - \sum_{i \in \mathcal{N}} V_i(p_i) - \boldsymbol{\zeta}^T \mathbf{q}(\mathbf{y}, \mathbf{p}(\mathbf{h})).
\tag{15}
$$

The dual function and the dual problem are, respectively,

$$
\varrho(\boldsymbol{\zeta}) := \max_{(\mathbf{y}, \mathbf{p}(\mathbf{h})) \in \mathcal{B}} \mathcal{L}(\mathbf{y}, \mathbf{p}(\mathbf{h}), \boldsymbol{\zeta})
\tag{16}
$$

$$
\mathsf{D} = \min_{\boldsymbol{\zeta} \ge \mathbf{0}} \varrho(\boldsymbol{\zeta}).
\tag{17}
$$

Since (14e) may be a non-convex constraint, the duality gap is in general, non-zero; i.e., $\mathsf{D} \ge \mathsf{P}$. Thus, solving (17) yields an upper bound on the optimal value P of (14). In the present formulation however, we have the following interesting result.

Proposition 1. *If the fading is continuous, then the duality gap is exactly zero, i.e.,* $\mathsf{P} = \mathsf{D}$.

A generalized version of Proposition 1, including a formal definition of continuous fading, is provided in Appendix A and connections to relevant results are made. The essential reason behind this strong duality is that the set of ergodic capacities resulting from all feasible power allocations is convex.

The requirement of continuous fading channels is not limiting since it holds for all practical fading models, such as Rayleigh, Rice, or Nakagami-m. Recall though that the dual problem is always convex. The subgradient method has traditionally been used to approximately solve (17), and also provide an intuitive layering interpretation of the network optimization problem [8]. The zero duality gap result is remarkable in the sense that it renders this layering optimal.

A corresponding result for unicast routing in uncoded networks has been proved in [22]. The fact that it holds for coded

networks with broadcast links, allows optimal integration of the network coding operations in the wireless protocol stack. The next subsection deals with this subject.

B. Subgradient Algorithm and Layer Separability

The dual problem (17) can in general be solved using the subgradient iterations [27, Sec. 8.2] indexed by ℓ

$$(\mathbf{y}(\ell), \mathbf{p}(\mathbf{h}; \ell)) \in \underset{(\mathbf{y},\mathbf{p}(\mathbf{h}))\in\mathcal{B}}{\arg\max} \mathcal{L}(\mathbf{y}, \mathbf{p}(\mathbf{h}), \boldsymbol{\zeta}(\ell)) \qquad (18a)$$

$$\boldsymbol{\zeta}(\ell+1) = [\boldsymbol{\zeta}(\ell) + \epsilon \mathbf{q}(\mathbf{y}(\ell), \mathbf{p}(\mathbf{h}; \ell))]^+ \qquad (18b)$$

where ϵ is a positive constant stepsize, and $[.]^+$ denotes projection onto the nonnegative orthant. The inclusion symbol ($\in$) allows for potentially multiple maxima. In (18b), $\mathbf{q}(\mathbf{y}(\ell), \mathbf{p}(\mathbf{h}; \ell))$ is a subgradient of the dual function $\varrho(\boldsymbol{\zeta})$ in (16) at $\boldsymbol{\zeta}(\ell)$. Next, we discuss the operations in (18) in detail.

For the Lagrangian obtained from (15), the maximization in (18a) can be separated into the following subproblems

$$a_i^m(\ell) \in \underset{a_{\min}^m \le a^m \le a_{\max}^m}{\arg\max} \left[U_m(a^m) - \sum_{t\in T^m} \nu_{s_m}^{mt}(\ell) a^m \right] \qquad (19a)$$

$$z_{iJ}^m(\ell) \in \underset{0 \le z_{iJ}^m \le z_{iJ}^{\max}}{\arg\max} \left[\sum_{\substack{K\subset N(i) \\ K\cap J\neq\emptyset}} \sum_{t\in T^m} \eta_{iK}^{mt}(\ell) - \xi_{iJ}(\ell) \right] z_{iJ}^m \qquad (19b)$$

$$x_{ij}^{mt}(\ell) \in \underset{0 \le x_{ij}^{mt} \le x_{ij}^{\max}}{\arg\max}$$

$$\left[\nu_i^{mt}(\ell)\mathbb{1}_{i\neq t} - \nu_j^{mt}(\ell)\mathbb{1}_{j\neq t} - \sum_{\substack{K\subset N(i) \\ j\in K}} \eta_{iK}^{mt}(\ell) \right] x_{ij}^{mt} \qquad (19c)$$

$$c_{iJ}(\ell) \in \underset{0 \le c_{iJ} \le c_{iJ}^{\max}}{\arg\max} [\xi_{iJ}(\ell) - \lambda_{iJ}(\ell)] c_{iJ} \qquad (19d)$$

$$p_i(\ell) \in \underset{0 \le p_i \le p_i^{\max}}{\arg\max} [\mu_i(\ell) p_i - V_i(p_i)] \qquad (19e)$$

$$\mathbf{p}(\mathbf{h}; \ell) \in \underset{\mathbf{p}(\mathbf{h})\in\Pi}{\arg\max} \sum_{f,(i,J)\in\mathcal{A}} \gamma_{iJ}^f(\mathbf{p}(\mathbf{h}), \mathbf{h}, \boldsymbol{\zeta}) \qquad (19f)$$

where

$$\gamma_{iJ}^f(\mathbf{p}(\mathbf{h}), \mathbf{h}, \boldsymbol{\zeta}) := \lambda_{iJ} C_{iJ}^f(\mathbf{p}(\mathbf{h}), \mathbf{h}) - \mu_i p_{iJ}^f(\mathbf{h}) \qquad (19g)$$

and $\mathbb{1}_X$ is the indicator function, which equals one if the expression X is true, and zero otherwise.

The physical layer subproblem (19f) implies per-fading state separability. Specifically, instead of optimizing over the class of power control policies, (19f) allows solving for the optimal power allocation for each fading state; that is,

$$\mathsf{P}(\mathbf{p}(\mathbf{h})) = \underset{\mathbf{p}(\mathbf{h})\in\Pi}{\max} \mathbb{E} \left[\sum_{f,(i,J)\in\mathcal{A}} \gamma_{iJ}^f(\mathbf{p}(\mathbf{h}), \mathbf{h}, \boldsymbol{\zeta}) \right]$$

$$= \mathbb{E} \left[\underset{\mathbf{p}(\mathbf{h})\in\Pi}{\max} \sum_{f,(i,J)\in\mathcal{A}} \gamma_{iJ}^f(\mathbf{p}(\mathbf{h}), \mathbf{h}, \boldsymbol{\zeta}) \right]. \qquad (20)$$

Note that problems (19a)–(19e) are convex and admit efficient solutions. The per-fading state power allocation subproblem (19f) however, may not necessarily be convex. For example, under the conflict graph model (cf. Example 1), the number of feasible power allocations may be exponential in the number of nodes. Finding an allocation that maximizes the objective function in (20) is equivalent to the NP-hard maximum weighted hyperarc matching problem [13]. Similarly, the capacity function and hence the objective function for the SINR model (cf. Example 2) is non-convex in general, and may be difficult to optimize.

This separable structure allows a useful layered interpretation of the problem. In particular, the transport layer subproblem (19a) gives the optimal exogenous rates allowed into the network; the network flow sub-problem (19b) yields the endogenous flow rates of coded packets on the hyperarcs; and the virtual flow sub-problem (19c) is responsible for determining the virtual flow rates between nodes and therefore the network code design. Likewise, the capacity sub-problem (19d) yields the link capacities and the power sub-problem (19e) provides the power control at the data link layer.

The layered architecture described so far also allows for optimal integration of network coding into the protocol stack. Specifically, the broadcast and virtual flows optimized respectively in (19b) and (19c), allow performing the combined routing-plus-network coding task at network layer. An implementation such as the one in [10] typically requires queues for both broadcast as well as virtual flows to be maintained here.

Next, the subgradient updates of (18b) become

$$\nu_i^{mt}(\ell+1) = [\nu_i^{mt}(\ell) + \epsilon \check{q}_\nu^{imt}(\ell)]^+ \qquad (21a)$$

$$\eta_{iK}^{mt}(\ell+1) = [\eta_{iK}^{mt}(\ell) + \epsilon \check{q}_\eta^{iKmt}(\ell)]^+ \qquad (21b)$$

$$\xi_{iJ}(\ell+1) = [\xi_{iJ}(\ell) + \epsilon \check{q}_\xi^{iJ}(\ell)]^+ \qquad (21c)$$

$$\lambda_{iJ}(\ell+1) = [\lambda_{iJ}(\ell) + \epsilon \check{q}_\lambda^{iJ}(\ell)]^+ \qquad (21d)$$

$$\mu_i(\ell+1) = [\mu_i(\ell) + \epsilon \check{q}_\mu^i(\ell)]^+ \qquad (21e)$$

where $\check{q}(\ell)$ are the subgradients at index ℓ given by

$$\check{q}_\nu^{imt}(\ell) = \sigma_i^m(\ell) + \sum_{(i,j)\in\mathcal{G}} x_{ji}^{mt}(\ell) - \sum_{(j,i)\in\mathcal{G}} x_{ij}^{mt}(\ell) \qquad (22a)$$

$$\check{q}_\eta^{iKmt}(\ell) = \sum_{j\in K} x_{ij}^{mt}(\ell) - \sum_{\substack{J:(i,J)\in\mathcal{A} \\ J\cap K\neq\emptyset}} z_{iJ}^m(\ell) \qquad (22b)$$

$$\check{q}_\xi^{iJ}(\ell) = \sum_m z_{iJ}^m(\ell) - c_{iJ}(\ell) \qquad (22c)$$

$$\check{q}_\lambda^{iJ}(\ell) = c_{iJ}(\ell) - \mathbb{E} \left[\sum_f C_{iJ}^f(\mathbf{p}(\mathbf{h}; \ell), \mathbf{h}) \right] \qquad (22d)$$

$$\check{q}_\mu^i(\ell) = \mathbb{E} \left[\sum_f \sum_{J:(i,J)\in\mathcal{A}} p_{iJ}^f(\mathbf{h}; \ell) \right] - p_i(\ell). \qquad (22e)$$

The physical layer updates (21d) and (21e) are again complicated since they involve the $\mathbb{E}[.]$ operations of (22d) and (22e). These expectations can be acquired via Monte Carlo simulations by solving (19f) for realizations of $\mathbf{h}$ and averaging over them. These realizations can be independently drawn from the distribution of $\mathbf{h}$, or they can be actual channel measurements. In fact, the latter is implemented in Section IV on the fly during network operation.

C. Convergence Results

This subsection provides convergence results for the subgradient iterations (18). Since the primal variables $(\mathbf{y}, \mathbf{p}(\mathbf{h}))$ and the capacity function $C_{iJ}^f(.)$ are bounded, it is possible to define an upper bound G on the subgradient norm; i.e., $\|\mathbf{q}(\mathbf{y}(\ell), \mathbf{p}(\mathbf{h}; \ell))\| \leq G$ for all $\ell \geq 1$.

Proposition 2. *For the subgradient iterations in* (19) *and* (21), *the best dual value converges to* D *upto a constant; i.e.,*

$$\lim_{s \to \infty} \min_{1 \leq \ell \leq s} \varrho(\boldsymbol{\zeta}(\ell)) \leq \mathsf{D} + \frac{\epsilon G^2}{2}. \quad (23)$$

This result is well known for dual (hence, convex) problems [27, Prop. 8.2.3]. However, the presence of an infinite-dimensional variable $\mathbf{p}(\mathbf{h})$ is a subtlety here. A similar case is dealt with in [22] and Proposition 2 follows from the results there.

Note that in the subgradient method (18), the sequence of primal iterates $\{\mathbf{y}(\ell)\}$ does not necessarily converge. However, a primal running average scheme can be used for finding the optimal primal variables $\mathbf{y}^*$ as summarized next. Recall that $f(\mathbf{y})$ denotes the objective function $\sum_m U_m(a^m) - \sum_i V_i(p_i)$.

Proposition 3. *For the running average of primal iterates*

$$\bar{\mathbf{y}}(s) := \frac{1}{s} \sum_{\ell=1}^{s} \mathbf{y}(\ell). \quad (24)$$

the following results hold:

  a) *There exists a sequence* $\{\mathring{\mathbf{p}}(\mathbf{h}; s)\}$ *such that* $(\bar{\mathbf{y}}(s), \mathring{\mathbf{p}}(\mathbf{h}; s)) \in \mathcal{B}$, *and also*

$$\lim_{s \to \infty} \left\| [\mathbf{q}(\bar{\mathbf{y}}(s), \mathring{\mathbf{p}}(\mathbf{h}; s))]^+ \right\| = 0. \quad (25)$$

  b) *The sequence* $f(\bar{\mathbf{y}}(s))$ *converges in the sense that*

$$\liminf_{s \to \infty} f(\bar{\mathbf{y}}(s)) \geq \mathsf{P} - \frac{\epsilon G^2}{2} \quad (26a)$$
$$and \quad \limsup_{s \to \infty} f(\bar{\mathbf{y}}(s)) \leq \mathsf{P}. \quad (26b)$$

Equation (25) asserts that the sequence $\{\bar{\mathbf{y}}(\ell)\}$ together with an associated $\{\mathring{\mathbf{p}}(\mathbf{h}; \ell)\}$ becomes asymptotically feasible. Moreover, (26) explicates the asymptotic suboptimality as a function of the stepsize, and the bound on the subgradient norm. Proposition 3 however, does not provide a way to actually find $\{\mathring{\mathbf{p}}(\mathbf{h}; \ell)\}$.

Averaging of the primal iterates is a well-appreciated method to obtain optimal primal solutions from dual subgradient methods in convex optimization [25]. Note though that the primal problem at hand is non-convex in general. Results related to Proposition 3 are shown in [23]. Proposition 3 follows in this paper as a special case result for a more general algorithm allowing for asynchronous subgradients and suitable for online network control, elaborated next.

IV. SUBGRADIENT ALGORITHM FOR NETWORK CONTROL

The algorithm in Section III-B finds the optimal operating point of (14) in an offline fashion. In the present section, the subgradient method is adapted so that it can be used for resource allocation during network operation.

The algorithm is motivated by Proposition 3 as follows. The exogenous arrival rates $a^m(\ell)$ generated by the subgradient

method [cf. (19a)] can be used as the instantaneous rate of the traffic admitted at the transport layer at time ℓ. Then, Proposition 3 guarantees that the long-term average transport layer rates will be optimal. Similar observations can be made for other rates in the network.

More generally, an online algorithm with the following characteristics is desirable.

- Time is divided in slots and each subgradient iteration takes one time slot. The channel is assumed to remain invariant per slot but is allowed to vary across slots.
- Each layer maintains its set of dual variables, which are updated according to (21) with a constant stepsize ϵ.
- The instantaneous transmission and reception rates at the various layers are set equal to the primal iterates at that time slot, found using (19).
- Proposition 3 ensures that the long-term average rates are optimal.

For network resource allocation problems such as those described in [5], the subgradient method naturally lends itself to an online algorithm with the aforementioned properties. This approach however cannot be directly extended to the present case because the dual updates (21d)–(21e) require an expectation operation, which needs prior knowledge of the exact channel distribution function for generation of independent realizations of $\mathbf{h}$ per time slot. Furthermore, although Proposition 3 guarantees the existence of a sequence of feasible power variables $\mathring{\mathbf{p}}(\mathbf{h}; s)$, it is not clear if one could find them since the corresponding running averages do not necessarily converge.

Towards adapting the subgradient method for network control, recall that the subgradients $\check{q}_\lambda^{iJ}$ and $\check{q}_\mu^i$ involve the following summands that require the expectation operations [cf. (22d) and (22e)]

$$\tilde{C}_{iJ}(\ell) := \mathbb{E}\left[\sum_f C_{iJ}^f(\mathbf{p}(\mathbf{h}; \ell), \mathbf{h}) \right] \quad (27)$$

$$\tilde{P}_i(\ell) := \mathbb{E}\left[\sum_{f, J:(i,J) \in \mathcal{A}} p_{iJ}^f(\mathbf{h}; \ell) \right]. \quad (28)$$

These expectations can however be approximated by averaging over actual channel realizations. To do this, the power allocation subproblem (19f) must be solved repeatedly for a prescribed number of time slots, say S, while using the same Lagrange multipliers. This would then allow approximation of the $\mathbb{E}$ operations in (27) and (28) with averaging operations, performed over channel realizations at these time slots.

It is evident however, that the averaging operation not only consumes S time slots but also that the resulting subgradient is always outdated. Specifically, if the current time slot is of the form $\ell = KS + 1$ with $K = 0, 1, 2, \ldots$, the most recent approximations of $\tilde{C}_{iJ}$ and $\tilde{P}_i$ available are

$$\hat{C}_{iJ}(\ell - S) = \frac{1}{S} \sum_{\kappa=\ell-S}^{\ell-1} \sum_f C_{iJ}^f(\mathbf{p}(\mathbf{h}_\kappa; \ell - S), \mathbf{h}_\kappa) \quad (29a)$$

$$\hat{P}_i(\ell - S) = \frac{1}{S} \sum_{\kappa=\ell-S}^{\ell-1} \sum_{f, J:(i,J) \in \mathcal{A}} p_{iJ}^f(\mathbf{h}_\kappa; \ell - S). \quad (29b)$$

Algorithm 1: Asynchronous Subgradient Algorithm

---

1 Initialize $\zeta(1) = 0$ and $\hat{C}_{iJ}(1) = \hat{P}_i(1) = 0$.
Let N be the maximum number of subgradient iterations.

2 **for** $\ell = 1, 2, \ldots, N$, **do**

3   Calculate primal iterates $a^m(\ell)$, $x_{ij}^{mt}(\ell)$, $z_{iJ}^m(\ell)$, $c_{iJ}(\ell)$, and $p_i(\ell)$ [cf. (19a)-(19e)].

4   Calculate the optimal power allocation $\mathbf{p}(\mathbf{h}_\ell; \tau(\ell))$ by solving (19f) using $\mathbf{h}_\ell$ and $\zeta(\tau(\ell))$.

5   Update dual iterates $\nu_i^{mt}(\ell+1)$, $\eta_{ik}^{mt}(\ell+1)$ and $\xi_{ij}(\ell+1)$ from the current primal iterates evaluated in Line 3 [cf. (21a)-(21c)].

6   **if** $\ell - \tau(\ell) = S$, **then**

7   Calculate $\hat{C}_{iJ}(\tau(\ell))$ and $\hat{P}_i(\tau(\ell))$ as in (29).

8   **end**

9   Update the dual iterates $\lambda_{iJ}(\ell+1)$ and $\mu_i(\ell+1)$:

$$\lambda_{iJ}(\ell+1) = \left[\lambda_{iJ}(\ell) + \epsilon(c_{iJ}(\ell) - \hat{C}_{iJ}(\tau(\ell))) \right]^+$$

$$\mu_i(\ell+1) = \left[\mu_i(\ell) + \epsilon(\hat{P}_i(\tau(\ell)) - p_i(\ell)) \right]^+.$$

10   Network Control: Use the current iterates $a^m(\ell)$ for flow control; $x_{ij}^{mt}(\ell)$ and $z_{iJ}^m(\ell)$ for routing and network coding; $c_{iJ}(\ell)$ for link rate control; and $\mathbf{p}(\mathbf{h}_\ell; \tau(\ell))$ for instantaneous power allocation.

11 **end**

Here, the power allocations are calculated using (19f) with the *old* multipliers $\lambda_{iJ}(\ell - S)$ and $\mu_i(\ell - S)$. The presence of outdated subgradient summands motivates the use of an asynchronous subgradient method such as the one in [24].

Specifically, the dual updates still occur at every time slot but are allowed to use subgradients with outdated summands. Thus, $\hat{C}_{iJ}(\ell - S)$ and $\hat{P}_i(\ell - S)$ are used instead of the corresponding $\mathbb{E}[.]$ terms in (22d) and (22e) at the current time ℓ. Further, since the averaging operation consumes another S time slots, the same summands are also used for times $\ell + 1$, $\ell + 2, \ldots, \ell + S - 1$. At time $\ell + S$, power allocations from the time slots ℓ, $\ell + 1$, $\ell + S - 1$ become available, and are used for calculating $\hat{C}_{iJ}(\ell)$ and $\hat{P}_i(\ell)$, which then serve as the more recent subgradient summands. Note that a subgradient summand such as $\hat{C}_{iJ}$ is at least S and at most $2S - 1$ slots old.

The asynchronous subgradient method is summarized as Algorithm 1. The algorithm uses the function $\tau(\ell)$ which outputs the time of most recent averaging operation, that is,

$$\tau(\ell) = \max\{S\lfloor(\ell - S - 1)/S\rfloor + 1, 1\} \quad \forall \ \ell \geq 1. \quad (31)$$

Note that $S \leq \ell - \tau(\ell) \leq 2S - 1$. Recall also that the subgradient components $\hat{C}_{iJ}$ and $\hat{P}_i$ are evaluated only at times $\tau(\ell)$.

The following proposition gives the dual convergence result on this algorithm. Define $\bar{G}$ as the bound $\left\| [\hat{\mathbf{C}}^T \ \hat{\mathbf{P}}^T]^T \right\| \leq \bar{G}$ where $\hat{\mathbf{C}}$ and $\hat{\mathbf{P}}$ are formed by stacking the terms $\mathbb{E}\left[\sum_f C_{iJ}^f(\mathbf{p}(\mathbf{h}), \mathbf{h}) \right]$ and $\mathbb{E}\left[\sum_{f,J} p_{iJ}^f(\mathbf{h}) \right]$, respectively.

Proposition 4. *If the maximum delay of the asynchronous counterparts of physical layer updates (21d) and (21e) is D, then:*

a) *The sequence of dual iterates $\{\zeta(\ell)\}$ is bounded; and*

b) *The best dual value converges to D up to a constant:*

$$\lim_{s \to \infty} \min_{1 \leq \ell \leq s} \varrho(\zeta(\ell)) \leq \mathsf{D} + \frac{\epsilon G^2}{2} + 2\epsilon D\bar{G}G. \quad (32)$$

Thus, the suboptimality in the asynchronous subgradient over the synchronous version is bounded by a constant proportional to $D = 2S - 1$. Consequently, the asynchronous subgradient might need a smaller stepsize (and hence, more iterations) to reach a given distance from the optimal.

The convergence of asynchronous subgradient methods for convex problems such as (17) has been studied in [24, Sec. 6] for a diminishing stepsize. Proposition 4 provides a complementary result for constant stepsizes.

Again, as with the synchronous version, the primal running averages also converge to within a constant from the optimal value of (14). This is stated formally in the next proposition.

Proposition 5. *If the maximum delay of the asynchronous counterparts of physical layer updates (21d) and (21e) is D, then:*

a) *There exists a sequence $\mathring{\mathbf{p}}(\mathbf{h}; s)$ such that $(\bar{\mathbf{y}}(s), \mathring{\mathbf{p}}(\mathbf{h}; s)) \in \mathcal{B}$ and*

$$\lim_{s \to \infty} \left\| [\mathbf{q}(\bar{\mathbf{y}}(s), \mathring{\mathbf{p}}(\mathbf{h}; s))]^+ \right\| = 0. \quad (33)$$

b) *The sequence $f(\bar{\mathbf{y}}(s))$ converges in the following sense:*

$$\liminf_{s \to \infty} f(\bar{\mathbf{y}}(s)) \geq \mathsf{P} - \frac{\epsilon G^2}{2} - 2\epsilon D\bar{G}G \quad (34a)$$

$$and \quad \limsup_{s \to \infty} f(\bar{\mathbf{y}}(s)) \leq \mathsf{P}. \quad (34b)$$

Note that as with the synchronous subgradient, the primal running averages are still asymptotically feasible, but the bound on their suboptimality increases by a term proportional to the delay D in the physical layer updates. Of course, all the results in Propositions 4 and 5 reduce to the corresponding results in Propositions 2 and 3 on setting $D = 0$. Interestingly, there is no similar result for primal convergence in asynchronous subgradient methods even for convex problems.

Finally, the following remarks on the online nature of the algorithm and the implementation of the Lagrangian maximizations in (19) are in order.

Remark 1. Algorithm 1 has several characteristics of an online adaptive algorithm. In particular, prior knowledge of the channel distribution is not needed in order to run the algorithm since the expectation operations are replaced by averaging over channel realizations on the fly. Likewise, running averages need not be evaluated; Proposition 5 ensures that the corresponding long-term averages will be near-optimal. Further, if at some time the network topology changes and the algorithm keeps running, it would be equivalent to restarting the entire algorithm with the current state as initialization. The algorithm is adaptive in this sense.

Remark 2. Each of the maximization operations (19a)–(19e) is easy, because it involves a single variable, concave objective, box constraints, and locally available Lagrange multipliers. The power control subproblem (19f) however may be hard and require centralized computation in order to obtain a (near-) optimal solution. For the conflict graph model, see [13], [28] and references therein for a list of approximate
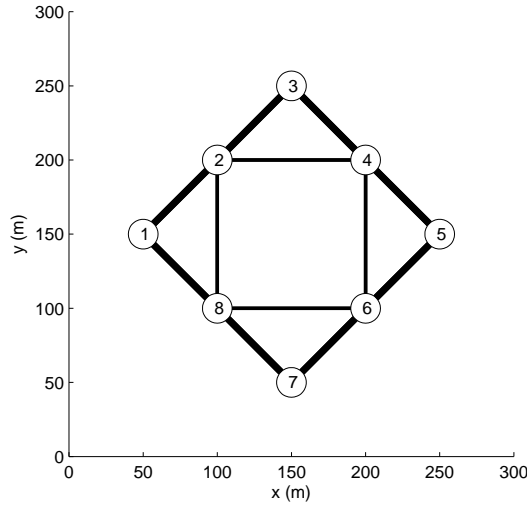
Fig. 1. The wireless network used in the simulations. The edges indicate the neighborhood of each node. The thickness of the edges is proportional to the mean of the corresponding channel.

Fig. 2. Evolution of the utility function $f(\bar{\mathbf{y}}(s))$ and best dual value $\rho_{\text{best}}(s) = \min_{\ell \leq s} \varrho(\boldsymbol{\zeta}(\ell))$ for $\epsilon = 0.15$ and $S = 50$.

TABLE I
SIMULATION PARAMETERS

F	2
h_{ij}^f	Exponential with mean $\bar{h}_{ij}^f = 0.1(d_{ij}/d_0)^{-2}$ for all $(i,j) \in \mathcal{G}$ and f, where $d_0 = 20$m and d_{ij} is the distance between the nodes i and j; links are reciprocal, i.e., $h_{ij}^f = h_{ji}^f$.
N_j	Noise power, evaluated using $d_{ij} = 100$m in the expression for $\bar{h}_{ij}^f$ above
$p_{\max}^f$	5 W/Hz for all f
$p_i^{\max}$	5 W/Hz for all $i \in \mathcal{N}$
$a_{\max}^m$	5 bps/Hz for all m
$a_{\min}^m$	10^{-4} bps/Hz for all m
$c_{iJ}^{\max}$	interference-free capacity obtained for each $j \in J$ via waterfilling under $\mathbb{E}\left[\sum_f p^f(h_{ij}^f)\right] \leq p_i^{\max}$ for all $i \in \mathcal{N}$
$z_{iJ}^{\max}$	$c_{iJ}^{\max}/2$ for all $(i,J) \in \mathcal{A}$
$x_{ij}^{\max}$	$z_{iJ}^{\max}/2$ for $j \in J$ and $i \in \mathcal{N}$
$U_m(a^m)$	$\ln(a^m)$ for all m
$V_i(p_i)$	$10p_i^2$ for all $i \in \mathcal{N}$

algorithms. For the SINR model, solutions of (19f) could be based on approximation techniques in power control for digital subscriber lines (DSL)—see e.g., [23] and references therein—and efficient message passing protocols as in [11].

V. NUMERICAL TESTS

The asynchronous algorithm developed in Section IV is simulated on the wireless network shown in Fig. 1. The network has 8 nodes placed on a 300m × 300m area. Hyperarcs originating from node i are denoted by $(i, J) \in \mathcal{A}$ where $J \in 2^{N(i)} \setminus \emptyset$ i.e., the power set of the neighbors of i excluding the empty set. For instance, hyperarcs originating from node 1 are $(1, \{2\})$, $(1, \{8\})$ and $(1, \{2,8\})$. The network supports the two multicast sessions $\mathcal{S}_1 = \{1, \{4,6\}\}$ and $\mathcal{S}_2 = \{4, \{1,7\}\}$. Table I lists the parameter values used in the simulation.

The conflict graph model of Example 1 with secondary interference constraints is used. In order to solve the power contr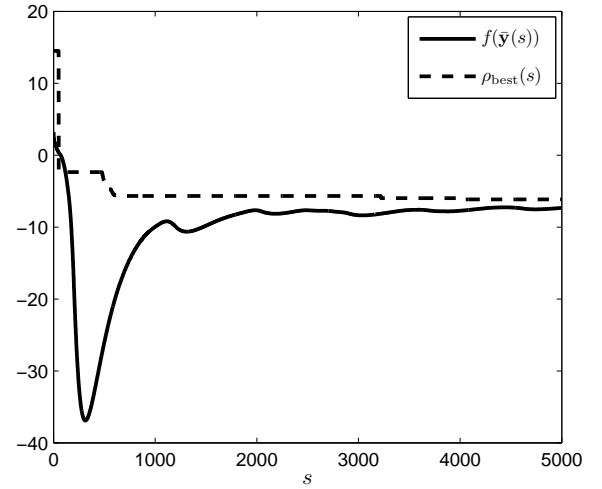ol subproblem (19f), we need to enumerate all possible sets of conflict free hyperarcs (cf. Example 1); these sets are called matchings. At each time slot, the aim is to find the matching that maximizes the objective function $\sum_{f,(i,J)} \gamma_{iJ}^f$. Note that since γ_{iJ}^f is a positive quantity, only maximal matchings, i.e., matchings with maximum possible cardinality, need to be considered. At each time slot, the following two steps are carried out.

S1) Find the optimal power allocation for each maximal matching. Note that the capacity of an active hyperarc is a function of the power allocation over that hyperarc alone [cf. (3) and (4)]. Thus, the maximization in (19f) can be solved separately for each hyperarc and tone. The resulting objective [cf. (19g)] is a concave function in a single variable, admitting an easy waterfilling-type solution.

S2) Evaluate the objective function (19f) for each maximal matching and for powers found in Step 2, and choose the matching with the highest resulting objective value.

It is well known that the enumeration of hyperarc matchings requires exponential complexity [13]. Since the problem at hand is small, full enumeration is used.

Fig. 2 shows the evolution of the utility function $f(\bar{\mathbf{y}}(s))$ and the best dual value up to the current iteration. The utility function is evaluated using the running average of the primal iterates [cf. (24)]. It can be seen that after a certain number of iterations, the primal and dual values remain very close corroborating the vanishing duality gap.

Fig. 3 shows the evolution of the utility function for different values of S. Again the utility function converges to a near-optimal value after sufficient number of iterations. Note however that the gap from the optimal dual value increases for large values of S, such as $S = 60$ (cf. Proposition 5).

Finally, Fig. 4 shows the optimal values of certain optimization variables. Specifically the two subplots show all the virtual flows to given sinks for each of the multicast sessions, namely, $\{s^1 = 1, t = 6\}$ and $\{s^2 = 4, t = 7\}$, respectively. The thickness and the gray level of the edges is proportional to the magnitude of the virtual flows. It can be observed that most virtual flows are concentrated along the shorter paths between the source and the sink. Also, the radius of the circles
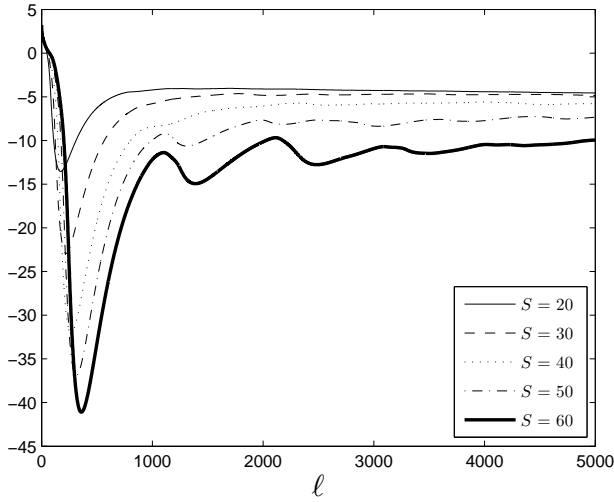
Fig. 3. Evolution of the utility function $f(\bar{\mathbf{y}}(s))$ for different values of S with stepsize $\epsilon = 0.15$.

representing the nodes is proportional to the optimal average power consumption. It can be seen that the inner nodes 2, 4, 6, and 8 consume more power than the outer ones, 1, 3, 5, and 7. This is because the inner nodes have more neighbors, and thus more opportunities to transmit. Moreover, the outer nodes are all close to their neighbors.

VI. CONCLUSIONS

This paper formulates a cross-layer optimization problem for multicast networks where nodes perform intra-session network coding, and operate over fading broadcast links. Zero duality gap is established, rendering layered architectures optimal.

Leveraging this result, an adaptation of the subgradient method suitable for network control is also developed. The method is asynchronous, because the physical layer returns its contribution to the subgradient vector with delay. Using the subgradient vector, primal iterates in turn dictate routing, network coding, and resource allocation. It is established that network variables, such as the long-term average rates admitted into the network layer, converge to near-optimal values, and the suboptimality bound is provided explicitly as a function of the delay in the subgradient evaluation.

APPENDIX A
STRONG DUALITY FOR THE NETWORKING PROBLEM (14)

This appendix formulates a general version of problem (14), and gives results about its duality gap. Let $\mathbf{h}$ be the random channel vector in $\Omega := \mathbb{R}_+^{d_\mathbf{h}}$, where $\mathbb{R}_+$ denotes the nonnegative reals, and $d_\mathbf{h}$ the dimensionality of $\mathbf{h}$. Let $\mathcal{D}$ be the σ-field of Borel sets in Ω, and $P_\mathbf{h}$ the distribution of $\mathbf{h}$, which is a probability measure on $\mathcal{D}$.

As in (14), consider two optimization variables: the vector $\mathbf{y}$ constrained to a subset $\mathcal{B}_\mathbf{y}$ of the Euclidean space $\mathbb{R}^{d_\mathbf{y}}$; and the function $\mathbf{p} : \Omega \to \mathbb{R}^{d_\mathbf{p}}$ belonging to an appropriate set of functions $\mathcal{P}$. In the networking problem, the aforementioned function is the power allocation $\mathbf{p}(\mathbf{h})$, and set $\mathcal{P}$ consists of the power allocation functions satisfying instantaneous constraints, such as spectral mask or hyperarc
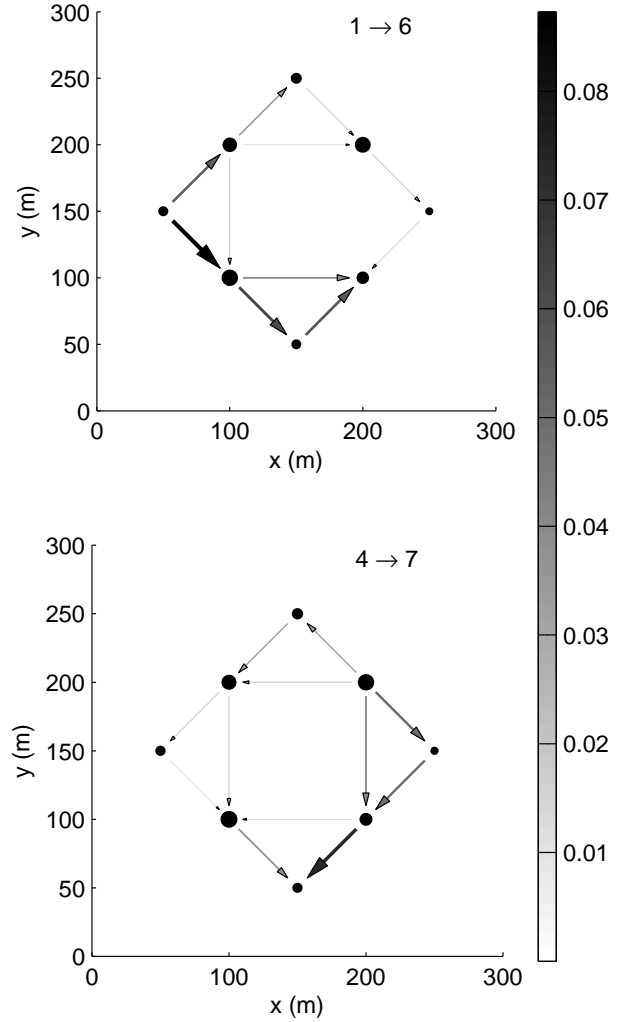
Fig. 4. Some of the optimal primal values after 5000 iterations with $\epsilon = 0.15$ and $S = 40$. The gray level of the edges corresponds to values of virtual flows according to the color bar on the right, with units bps/Hz.

scheduling constraints (cf. also Examples 1 and 2). Henceforth, the function variable will be denoted by $\mathbf{p}$ instead of $\mathbf{p}(\mathbf{h})$, for brevity. Let Π be a subset of $\mathbb{R}^{d_\mathbf{p}}$. Then $\mathcal{P}$ is defined as the set of functions taking values in Π.

$$\mathcal{P} := \{\mathbf{p} \text{ measurable} \mid \mathbf{p}(\mathbf{h}) \in \Pi \text{ for almost all } \mathbf{h} \in \Omega\}. \quad (35)$$

The network optimization problem (14) can be written in the general form

$$\mathsf{P} = \max \quad f(\mathbf{y}) \quad (36a)$$
$$\text{subj. to} \quad \mathbf{g}(\mathbf{y}) + \mathbb{E}[\mathbf{v}(\mathbf{p}(\mathbf{h}), \mathbf{h})] \leq \mathbf{0} \quad (36b)$$
$$\mathbf{y} \in \mathcal{B}_\mathbf{y}, \ \mathbf{p} \in \mathcal{P} \quad (36c)$$

where $\mathbf{g}$ and $\mathbf{v}$ are $\mathbb{R}^d$-valued functions describing d constraints. The formulation also subsumes similar problems in the unicast routing framework such as those in [22], [23].

Evidently, problem (14) is a special case of (36). If inequalities (14b)–(14f) are rearranged to have zeros on the right hand side, function $\mathbf{v}(\mathbf{p}(\mathbf{h}), \mathbf{h})$ will simply have zeros in the entries that correspond to constraints (14b)–(14d). The function $\mathbf{q}(\mathbf{y}, \mathbf{p}(\mathbf{h}))$ defined before (15) equals $\mathbf{g}(\mathbf{y}) + \mathbb{E}[\mathbf{v}(\mathbf{p}(\mathbf{h}), \mathbf{h})]$.

The following assumptions regarding (36) are made:

AS1. Constraint set $\mathcal{B}_\mathbf{y}$ is convex, closed, bounded, and in the interior of the domains of functions $f(\mathbf{y})$ and $\mathbf{g}(\mathbf{y})$. Set Π is closed, bounded, and in the interior of the domain of function $\mathbf{v}(.,\mathbf{h})$ for all $\mathbf{h}$.

AS2. Function $f(\cdot)$ is concave, $\mathbf{g}(\cdot)$ is convex, and $\mathbf{v}(\mathbf{p}(\mathbf{h}),\mathbf{h})$ is integrable whenever $\mathbf{p}$ is measurable. Furthermore, there is a $\bar{G} > 0$ such that $\|\mathbb{E}[\mathbf{v}(\mathbf{p}(\mathbf{h}),\mathbf{h})]\| \leq \bar{G}$, whenever $\mathbf{p} \in \mathcal{P}$.

AS3. Random vector $\mathbf{h}$ is continuous;[1] and

AS4. There exist $\mathbf{y}' \in \mathcal{B}_\mathbf{y}$ and $\mathbf{p}' \in \mathcal{P}$ such that (36b) holds as strict inequality (Slater constraint qualification).

Note that these assumptions are natural for the network optimization problem (14). Specifically, $\mathcal{B}_\mathbf{y}$ are the box constraints for variables a^m, x_{ij}^{mt}, z_{iJ}^m, c_{iJ}, and p_i; and Π gives the instantaneous power allocation constraints. The function $f(\mathbf{y})$ is selected concave and $\mathbf{g}(\mathbf{y})$ is linear. Moreover, the entries of $\mathbf{v}(\mathbf{p}(\mathbf{h}),\mathbf{h})$ corresponding to (14f) are bounded because the set Π is bounded. For the same reason, the ergodic capacities $\mathbb{E}[C_{iJ}^f(\mathbf{p}(\mathbf{h}),\mathbf{h})]$ are bounded.

While (36) is not convex in general, it is separable [29, Sec. 5.1.6]. The Lagrangian (keeping constraints (36c) implicit) and the dual function are, respectively [cf. also (15) and (16)]

$$\mathcal{L}(\mathbf{y},\mathbf{p},\boldsymbol{\zeta}) = f(\mathbf{y}) - \boldsymbol{\zeta}^T\Big(\mathbf{g}(\mathbf{y}) + \mathbb{E}[\mathbf{v}(\mathbf{p}(\mathbf{h}),\mathbf{h})]\Big) \quad (37)$$

$$\varrho(\boldsymbol{\zeta}) := \max_{\mathbf{y}\in\mathcal{B}_\mathbf{y},\,\mathbf{p}\in\mathcal{P}} \mathcal{L}(\mathbf{y},\mathbf{p},\boldsymbol{\zeta}) = \psi(\boldsymbol{\zeta}) + \phi(\boldsymbol{\zeta}). \quad (38)$$

where $\boldsymbol{\zeta}$ denotes the vector of Lagrange multipliers and

$$\psi(\boldsymbol{\zeta}) := \max_{\mathbf{y}\in\mathcal{B}_\mathbf{y}} \big\{f(\mathbf{y}) - \boldsymbol{\zeta}^T\mathbf{g}(\mathbf{y})\big\} \quad (39a)$$

$$\phi(\boldsymbol{\zeta}) := \max_{\mathbf{p}\in\mathcal{P}} \boldsymbol{\zeta}^T\mathbb{E}[\mathbf{v}(\mathbf{p}(\mathbf{h}),\mathbf{h})]. \quad (39b)$$

The additive form of the dual function is a consequence of the separable structure of the Lagrangian. Further, AS1 and AS2 ensure that the domain of $\varrho(\boldsymbol{\zeta})$ is $\mathbb{R}^d$. Finally, the dual problem becomes [cf. also (17)]

$$\mathsf{D} = \min_{\boldsymbol{\zeta}\geq\mathbf{0}} \varrho(\boldsymbol{\zeta}). \quad (40)$$

As $\mathbf{p}$ varies in $\mathcal{P}$, define the range of $\mathbb{E}[\mathbf{v}(\mathbf{p}(\mathbf{h}),\mathbf{h})]$ as

$$\mathcal{R} := \big\{\mathbf{w}\in\mathbb{R}^d \,|\, \mathbf{w} = \mathbb{E}[\mathbf{v}(\mathbf{p}(\mathbf{h}),\mathbf{h})] \text{ for some } \mathbf{p}\in\mathcal{P}\big\}. \quad (41)$$

The following lemma demonstrating the convexity of $\mathcal{R}$ plays a central role in establishing the zero duality gap of (36), and in the recovery of primal variables from the subgradient method.

Lemma 1. *If AS1-AS3 hold, then the set $\mathcal{R}$ is convex.*

The proof relies on Lyapunov's convexity theorem [30]. Recently, an extension of Lyapunov's theorem [30, Extension 1] has been applied to show zero duality gap of power control problems in DSL [26]. This extension however does not apply here, as indicated in the ensuing proof. In a related contribution [22], it is shown that the perturbation function of a problem similar to (36) is convex; the claim of Lemma 1 though is quite different.

---

[1]Formally, this is equivalent to saying that $P_\mathbf{h}$ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}_+^{d_\mathbf{h}}$. In more practical terms, it means that $\mathbf{h}$ has a probability density function without deltas.

Proof of Lemma 1: Let $\mathbf{r}_1$ and $\mathbf{r}_2$ denote arbitrary points in $\mathcal{R}$, and let $\alpha \in (0,1)$ be arbitrary. By the definition of $\mathcal{R}$, there are functions $\mathbf{p}_1$ and $\mathbf{p}_2$ in $\mathcal{P}$ such that

$$\mathbf{r}_1 = \int \mathbf{v}(\mathbf{p}_1(\mathbf{h}),\mathbf{h})dP_\mathbf{h} \text{ and } \mathbf{r}_2 = \int \mathbf{v}(\mathbf{p}_2(\mathbf{h}),\mathbf{h})dP_\mathbf{h}. \quad (42)$$

Now define

$$\mathbf{u}(E) := \begin{bmatrix} \int_E \mathbf{v}(\mathbf{p}_1(\mathbf{h}),\mathbf{h})dP_\mathbf{h} \\ \int_E \mathbf{v}(\mathbf{p}_2(\mathbf{h}),\mathbf{h})dP_\mathbf{h} \end{bmatrix}, \quad E \in \mathcal{D}. \quad (43)$$

The set function $\mathbf{u}(E)$ is a nonatomic vector measure on $\mathcal{D}$, because $P_\mathbf{h}$ is nonatomic (cf. AS3) and the functions $\mathbf{v}(\mathbf{p}_1(\mathbf{h}),\mathbf{h})$ and $\mathbf{v}(\mathbf{p}_2(\mathbf{h}),\mathbf{h})$ are integrable (cf. AS2); see [31] for definitions. Hence, Lyapunov's theorem applies to $\mathbf{u}(E)$; see also [30, Extension 1] and [22, Lemma 1].

Specifically, consider a null set Φ in $\mathcal{D}$, i.e., a set with $P_\mathbf{h}(\Phi) = 0$, and the whole space $\Omega \in \mathcal{D}$. It holds that $\mathbf{u}(\Phi) = \mathbf{0}$ and $\mathbf{u}(\Omega) = [\mathbf{r}_1^T, \mathbf{r}_2^T]^T$. For the chosen α, Lyapunov's theorem asserts that there exists a set $E_\alpha \in \mathcal{D}$ such that (E_α^c denotes the complement of E_α)

$$\mathbf{u}(E_\alpha) = \alpha\mathbf{u}(\Omega) + (1-\alpha)\mathbf{u}(\Phi) = \alpha\begin{bmatrix}\mathbf{r}_1\\\mathbf{r}_2\end{bmatrix} \quad (44a)$$

$$\mathbf{u}(E_\alpha^c) = \mathbf{u}(\Omega) - \mathbf{u}(E_\alpha) = (1-\alpha)\begin{bmatrix}\mathbf{r}_1\\\mathbf{r}_2\end{bmatrix}. \quad (44b)$$

Now using these E_α and E_α^c, define

$$\mathbf{p}_\alpha(\mathbf{h}) = \begin{cases} \mathbf{p}_1(\mathbf{h}), & \mathbf{h} \in E_\alpha \\ \mathbf{p}_2(\mathbf{h}), & \mathbf{h} \in E_\alpha^c. \end{cases} \quad (45)$$

It is easy to show that $\mathbf{p}_\alpha(\mathbf{h}) \in \mathcal{P}$. In particular, the function $\mathbf{p}_\alpha(\mathbf{h})$ can written as $\mathbf{p}_\alpha(\mathbf{h}) = \mathbf{p}_1(\mathbf{h})\mathbb{1}_{E_\alpha} + \mathbf{p}_2(\mathbf{h})\mathbb{1}_{E_\alpha^c}$, where $\mathbb{1}_E$ is the indicator function of a set $E \in \mathcal{D}$. Hence it is measurable, as sum of measurable functions. Moreover, we have that $\mathbf{p}_\alpha(\mathbf{h}) \in \Pi$ for almost all $\mathbf{h}$, because $\mathbf{p}_1(\mathbf{h})$ and $\mathbf{p}_2(\mathbf{h})$ satisfy this property. The need to show $\mathbf{p}_\alpha(\mathbf{h}) \in \mathcal{P}$ makes [30, Extension 1] not directly applicable here.

Thus, $\mathbf{p}_\alpha(\mathbf{h}) \in \mathcal{P}$ and satisfies [cf. (44)]

$$\int \mathbf{v}(\mathbf{p}_\alpha(\mathbf{h}),\mathbf{h})dP_\mathbf{h} = \int_{E_\alpha} \mathbf{v}(\mathbf{p}_1(\mathbf{h}),\mathbf{h})dP_\mathbf{h}$$
$$+ \int_{E_\alpha^c} \mathbf{v}(\mathbf{p}_2(\mathbf{h}),\mathbf{h})dP_\mathbf{h} = \alpha\mathbf{r}_1 + (1-\alpha)\mathbf{r}_2. \quad (46)$$

Therefore, $\alpha\mathbf{r}_1 + (1-\alpha)\mathbf{r}_2 \in \mathcal{R}$.　□

Finally, the zero duality gap result follows from Lemma 1 and is stated in the following proposition.

Proposition 6. *If AS1-AS4 hold, then problem* (36) *has zero duality gap, i.e.,* $\mathsf{P} = \mathsf{D}$. *Furthermore, the values* P *and* D *are finite, the dual problem* (40) *has an optimal solution, and the set of optimal solutions of* (40) *is bounded.*

Proof: Function $f(\mathbf{y})$ is continuous on $\mathcal{B}_\mathbf{y}$ since it is convex (cf. AS1 and AS2) [27, Prop. 1.4.6]. This, combined with the compactness of $\mathcal{B}_\mathbf{y}$, shows that the optimal primal value P is finite. Consider the set

$$\mathcal{W} := \big\{(w_1,\ldots,w_d,u) \in \mathbb{R}^{d+1} \mid f(\mathbf{y}) \leq u,$$

$$\mathbf{g}(\mathbf{y}) + \mathbb{E}[\mathbf{v}(\mathbf{p}(\mathbf{h}), \mathbf{h})] \leq \mathbf{w} \text{ for some } \mathbf{y} \in \mathcal{B}_{\mathbf{y}},\ \mathbf{p} \in \mathcal{P} \}. \tag{47}$$

Using Lemma 1, it is easy to verify that set $\mathcal{W}$ is convex. The rest of the proof follows that of [29, Prop. 5.3.1 and 5.1.4], using the finiteness of P and Slater constraint qualification (cf. AS4).

The boundedness of the optimal dual set is a standard result for convex problems under Slater constraint qualification and finiteness of optimal primal value; see e.g., [27, Prop. 6.4.3] and [25, p. 1762]. The proof holds also in the present setup since P is finite, P = D, and AS4 holds. $\square$

APPENDIX B
DUAL AND PRIMAL CONVERGENCE RESULTS

This appendix formulates the synchronous and asynchronous subgradient methods for the generic problem (36); and establishes the convergence claims in Propositions 2–5. Note that Propositions 2 and 3 follow from Propositions 4 and 5, respectively, upon setting the delay $D = 0$.

Starting from an arbitrary $\boldsymbol{\zeta}(1) \geq \mathbf{0}$, the subgradient iterations for (40) indexed by $\ell \in \mathbb{N}$ are [cf. also (18)]

$$\mathbf{y}(\ell) \in \arg\max_{\mathbf{y} \in \mathcal{B}_{\mathbf{y}}} \{ f(\mathbf{y}) - \boldsymbol{\zeta}^T(\ell) \mathbf{g}(\mathbf{y}) \} \tag{48a}$$

$$\mathbf{p}(.;\ell) \in \arg\max_{\mathbf{p} \in \mathcal{P}} \boldsymbol{\zeta}^T(\ell) \mathbb{E}[\mathbf{v}(\mathbf{p}(\mathbf{h}), \mathbf{h})] \tag{48b}$$

and $\qquad \boldsymbol{\zeta}(\ell+1) = [\boldsymbol{\zeta}(\ell) + \epsilon(\check{\boldsymbol{g}}(\ell) + \check{\boldsymbol{v}}(\ell))]^+ \tag{48c}$

where $\check{\boldsymbol{g}}(\ell)$ and $\check{\boldsymbol{v}}(\ell)$ are the subgradients of functions $\psi(\boldsymbol{\zeta})$ and $\phi(\boldsymbol{\zeta})$, defined as [cf. also (22)]

$$\check{\boldsymbol{g}}(\ell) := \mathbf{g}(\mathbf{y}(\ell)) \tag{49a}$$

$$\check{\boldsymbol{v}}(\ell) := \mathbb{E}[\mathbf{v}(\mathbf{p}(\mathbf{h};\ell), \mathbf{h})]. \tag{49b}$$

The iteration in (48c) is synchronous, because at every ℓ, both maximizations (48a) and (48b) are performed using the current Lagrange multiplier $\boldsymbol{\zeta}(\ell)$. An *asynchronous* method is also of interest and operates as follows. Here, the component $\check{\boldsymbol{v}}$ of the overall subgradient used at ℓ does not necessarily correspond to the Lagrange multiplier $\boldsymbol{\zeta}(\ell)$, but to the Lagrange multiplier at a time $\tau(\ell) \leq \ell$. Noting that the maximizer in (48b) is $\mathbf{p}(.;\tau(\ell)))$ and the corresponding subgradient component used at ℓ is $\check{\boldsymbol{v}}(\tau(\ell))$, the iteration takes the form

$$\boldsymbol{\zeta}(\ell+1) = [\boldsymbol{\zeta}(\ell) + \epsilon(\check{\boldsymbol{g}}(\ell) + \check{\boldsymbol{v}}(\tau(\ell)))]^+,\ \ell \in \mathbb{N}. \tag{50}$$

The difference $\ell - \tau(\ell)$ is the delay with which the subgradient component $\check{\boldsymbol{v}}$ becomes available. In Algorithm 1 for example, the delayed components are $\hat{C}_{iJ}(\tau(\ell))$ and $\hat{P}_i(\tau(\ell))$.

Next, we proceed to analyze the convergence of (50). Function $\mathbf{g}(\mathbf{y})$ is continuous on $\mathcal{B}_{\mathbf{y}}$ because it is convex [27, Prop. 1.4.6]. Then AS1 and AS2 imply that there exists a bound G such that for all $\mathbf{y} \in \mathcal{B}_{\mathbf{y}}$ and $\mathbf{p} \in \mathcal{P}$,

$$\|\mathbf{g}(\mathbf{y}) + \mathbb{E}[\mathbf{v}(\mathbf{p}(\mathbf{h}), \mathbf{h})]\| \leq G. \tag{51}$$

Due to this bound on the subgradient norm, algorithm (50) can be viewed as a special case of an approximate subgradient method [32]. We do not follow this line of analysis here though, because it does not take advantage of the source of the error in the subgradient—namely, that an old maximizer of the Lagrangian is used. Moreover, algorithm (50) can be viewed as a particular case of an ε-subgradient method (see [29, Sec. 6.3.2] for definitions). This connection is made in [24] which only deals with diminishing stepsizes; here results are proved for constant stepsizes. The following assumption is adopted for the delay $\ell - \tau(\ell)$.

AS5. There exists a finite $D \in \mathbb{N}$ such that $\ell - \tau(\ell) \leq D$ for all $\ell \in \mathbb{N}$.

AS5 holds for Algorithm 1 since the maximum delay there is $D = 2S - 1$. The following lemma collects the results needed for Propositions 2 and 4. Specifically, it characterizes the error term in the subgradient definition when $-\check{\boldsymbol{v}}(\tau(\ell))$ is used; and also relates successive iterates $\boldsymbol{\zeta}(\ell)$ and $\boldsymbol{\zeta}(\ell+1)$. The quantity $\bar{G}$ in the ensuing statement was defined in AS2.

Lemma 2. *Under AS1-AS5, the following hold for the sequence $\{\boldsymbol{\zeta}(\ell)\}$ generated by (50) for all $\boldsymbol{\theta} \geq \mathbf{0}$*

a) $-\check{\boldsymbol{v}}^T(\tau(\ell))(\boldsymbol{\theta} - \boldsymbol{\zeta}(\ell))$
$$\leq \phi(\boldsymbol{\theta}) - \phi(\boldsymbol{\zeta}(\ell)) + 2\epsilon DG\bar{G} \tag{52a}$$

b) $-(\check{\boldsymbol{g}}(\ell) + \check{\boldsymbol{v}}(\tau(\ell)))^T(\boldsymbol{\theta} - \boldsymbol{\zeta}(\ell))$
$$\leq \varrho(\boldsymbol{\theta}) - \varrho(\boldsymbol{\zeta}(\ell)) + 2\epsilon DG\bar{G} \tag{52b}$$

c) $\|\boldsymbol{\zeta}(\ell+1) - \boldsymbol{\theta}\|^2 - \|\boldsymbol{\zeta}(\ell) - \boldsymbol{\theta}\|^2$
$$\leq 2\epsilon[\varrho(\boldsymbol{\theta}) - \varrho(\boldsymbol{\zeta}(\ell))] + \epsilon^2 G^2 + 4\epsilon^2 DG\bar{G} \tag{52c}$$

Parts (a) and (b) of Lemma 2 assert that the vectors $-\check{\boldsymbol{v}}(\tau(\ell))$ and $-\check{\boldsymbol{g}}(\ell) - \check{\boldsymbol{v}}(\tau(\ell))$ are respectively ε-subgradients of $\phi(\boldsymbol{\zeta})$ and the dual function $\varrho(\boldsymbol{\zeta})$ at $\boldsymbol{\zeta}(\ell)$, with $\varepsilon = 2\epsilon DG\bar{G}$. Note that ε is a constant proportional to the delay D.

Proof of Lemma 2: a) The left-hand side of (52a) is

$$-\check{\boldsymbol{v}}^T(\tau(\ell))(\boldsymbol{\theta} - \boldsymbol{\zeta}(\ell)) = -\check{\boldsymbol{v}}^T(\tau(\ell))[\boldsymbol{\theta} - \boldsymbol{\zeta}(\tau(\ell))]$$
$$- \check{\boldsymbol{v}}^T(\tau(\ell))[\boldsymbol{\zeta}(\tau(\ell)) - \boldsymbol{\zeta}(\ell)]. \tag{53}$$

Applying the definition of the subgradient for $\phi(\boldsymbol{\zeta})$ at $\boldsymbol{\zeta}(\tau(\ell))$ to (53), it follows that

$$-\check{\boldsymbol{v}}^T(\tau(\ell))(\boldsymbol{\theta} - \boldsymbol{\zeta}(\ell)) \leq \phi(\boldsymbol{\theta}) - \phi(\boldsymbol{\zeta}(\tau(\ell)))$$
$$- \check{\boldsymbol{v}}^T(\tau(\ell))[\boldsymbol{\zeta}(\tau(\ell)) - \boldsymbol{\zeta}(\ell)]. \tag{54}$$

Now, adding and subtracting the same terms in the right-hand side of (54), we obtain

$$-\check{\boldsymbol{v}}^T(\tau(\ell))(\boldsymbol{\theta} - \boldsymbol{\zeta}(\ell)) \leq \phi(\boldsymbol{\theta}) - \phi(\boldsymbol{\zeta}(\ell))$$
$$+ \sum_{\kappa=1}^{\ell-\tau(\ell)} [\phi(\boldsymbol{\zeta}(\tau(\ell)+\kappa)) - \phi(\boldsymbol{\zeta}(\tau(\ell)+\kappa-1))]$$
$$- \sum_{\kappa=1}^{\ell-\tau(\ell)} \check{\boldsymbol{v}}^T(\tau(\ell))[\boldsymbol{\zeta}(\tau(\ell)+\kappa-1) - \boldsymbol{\zeta}(\tau(\ell)+\kappa)]. \tag{55}$$

Applying the definition of the subgradient for $\phi(\boldsymbol{\zeta})$ at $\boldsymbol{\zeta}(\tau(\ell)+\kappa)$ to (55), it follows that

$$-\check{\boldsymbol{v}}^T(\tau(\ell))(\boldsymbol{\theta} - \boldsymbol{\zeta}(\ell)) \leq \phi(\boldsymbol{\theta}) - \phi(\boldsymbol{\zeta}(\ell))$$
$$+ \sum_{\kappa=1}^{\ell-\tau(\ell)} \check{\boldsymbol{v}}^T(\tau(\ell)+\kappa)[\boldsymbol{\zeta}(\tau(\ell)+\kappa-1) - \boldsymbol{\zeta}(\tau(\ell)+\kappa)]$$
$$- \sum_{\kappa=1}^{\ell-\tau(\ell)} \check{\boldsymbol{v}}^T(\tau(\ell))[\boldsymbol{\zeta}(\tau(\ell)+\kappa-1) - \boldsymbol{\zeta}(\tau(\ell)+\kappa)]. \tag{56}$$

Using the Cauchy-Schwartz inqueality, (56) becomes

$$-\check{\boldsymbol{v}}^T(\tau(\ell))\left(\boldsymbol{\theta}-\boldsymbol{\zeta}(\ell)\right) \leq \phi(\boldsymbol{\theta}) - \phi(\boldsymbol{\zeta}(\ell))$$
$$+ \sum_{\kappa=1}^{\ell-\tau(\ell)} \left(\|\check{\boldsymbol{v}}(\tau(\ell)+\kappa)\| + \|\check{\boldsymbol{v}}(\tau(\ell))\|\right)$$
$$\cdot \|\boldsymbol{\zeta}(\tau(\ell)+\kappa-1) - \boldsymbol{\zeta}(\tau(\ell)+\kappa)\|. \quad (57)$$

Now, write the subgradient iteration [cf. (50)] at $\tau(\ell)+\kappa-1$:

$$\boldsymbol{\zeta}(\tau(\ell)+\kappa) = \big[\boldsymbol{\zeta}(\tau(\ell)+\kappa-1)$$
$$+ \epsilon\big(\check{\boldsymbol{g}}(\tau(\ell)+\kappa-1) + \check{\boldsymbol{v}}(\tau(\tau(\ell)+\kappa-1))\big)\big]^+. \quad (58)$$

Subtracting $\boldsymbol{\zeta}(\tau(\ell)+\kappa-1)$ from both sides of the latter and using the nonexpansive property of the projection [27, Prop. 2.2.1] followed by (51), one finds from (58) that

$$\|\boldsymbol{\zeta}(\tau(\ell)+\kappa) - \boldsymbol{\zeta}(\tau(\ell)+\kappa-1)\|$$
$$\leq \epsilon \|\check{\boldsymbol{g}}(\tau(\ell)+\kappa-1) + \check{\boldsymbol{v}}(\tau(\tau(\ell)+\kappa-1))\| \leq \epsilon G. \quad (59)$$

Finally, recall that $\|\check{\boldsymbol{v}}(\ell)\| \leq \bar{G}$ for all $\ell \in \mathbb{N}$ (cf. AS2), and $\ell - \tau(\ell) \leq D$ for all $\ell \in \mathbb{N}$ (cf. AS5). Applying the two aforementioned assumptions and (59) to (57), we obtain (52a).

b) This part follows readily from part a), using (38) and the definition of the subgradient for $\psi(\boldsymbol{\zeta})$ at $\boldsymbol{\zeta}(\ell)$ [cf. (49a)].

c) We have from (50) for all $\boldsymbol{\theta} \geq \boldsymbol{0}$ that

$$\|\boldsymbol{\zeta}(\ell+1) - \boldsymbol{\theta}\|^2 = \left\|[\boldsymbol{\zeta}(\ell) + \epsilon\left(\check{\boldsymbol{g}}(\ell) + \check{\boldsymbol{v}}(\tau(\ell))\right)]^+ - \boldsymbol{\theta}\right\|^2. \quad (60)$$

Due to the nonexpansive property of the projection, it follows that

$$\|\boldsymbol{\zeta}(\ell+1) - \boldsymbol{\theta}\|^2 \leq \|\boldsymbol{\zeta}(\ell) + \epsilon\left(\check{\boldsymbol{g}}(\ell) + \check{\boldsymbol{v}}(\tau(\ell))\right) - \boldsymbol{\theta}\|^2$$
$$= \|\boldsymbol{\zeta}(\ell) - \boldsymbol{\theta}\|^2 + \epsilon^2 \|\check{\boldsymbol{g}}(\ell) + \check{\boldsymbol{v}}(\tau(\ell))\|^2$$
$$+ 2\epsilon\left(\check{\boldsymbol{g}}(\ell) + \check{\boldsymbol{v}}(\tau(\ell))\right)^T \left(\boldsymbol{\zeta}(\ell) - \boldsymbol{\theta}\right). \quad (61)$$

Introducing (52b) and (51) into (61), (52c) follows. $\qquad\square$

The main convergence results for the synchronous and asynchronous subgradient methods are given by Propositions 2 and 4, respectively. Using Lemma 2, Proposition 4 is proved next.

Proof of Proposition 4: a) Let $\boldsymbol{\zeta}^*$ be an arbitrary dual solution. With $\mathbf{g}_i$ and $\mathbf{v}_i$ denoting the i-th entries of $\mathbf{g}$ and $\mathbf{v}$, respectively, define

$$\delta := \min_{1 \leq i \leq d} \left\{-\mathbf{g}_i(\mathbf{y}') - \mathbb{E}[\mathbf{v}_i(\mathbf{p}'(\mathbf{h}), \mathbf{h})]\right\} \quad (62)$$

where $\mathbf{y}'$ and $\mathbf{p}'$ are the strictly feasible variables in AS4. Note that $\delta > 0$ due to AS4.

We show that the following relation holds for all $\ell \geq 1$:

$$\|\boldsymbol{\zeta}(\ell) - \boldsymbol{\zeta}^*\| \leq \max\Big\{\|\boldsymbol{\zeta}(1) - \boldsymbol{\zeta}^*\|,$$
$$\frac{1}{\delta}(\mathsf{D} - f(\mathbf{y}')) + \frac{\epsilon G^2}{2\delta} + \frac{2\epsilon DG\bar{G}}{\delta} + \|\boldsymbol{\zeta}^*\| + \epsilon G\Big\}. \quad (63)$$

Eq. (63) implies that the sequence of Lagrange multipliers $\{\boldsymbol{\zeta}(\ell)\}$ is bounded, because the optimal dual set is bounded (cf. Proposition 6). Next, (63) is shown by induction.

It obviously holds for $\ell = 1$. Assume it holds for some $\ell \in \mathbb{N}$. It is proved next that it holds for $\ell+1$. Two cases are considered, depending on the value of $\varrho(\boldsymbol{\zeta}(\ell))$.

Case 1: $\varrho(\boldsymbol{\zeta}(\ell)) > \mathsf{D} + \epsilon G^2/2 + 2\epsilon DG\bar{G}$. Then eq. (52c) with $\boldsymbol{\theta} = \boldsymbol{\zeta}^*$ and $\varrho(\boldsymbol{\zeta}^*) = \mathsf{D}$ becomes

$$\|\boldsymbol{\zeta}(\ell+1) - \boldsymbol{\zeta}^*\|^2 \leq \|\boldsymbol{\zeta}(\ell) - \boldsymbol{\zeta}^*\|^2$$
$$- 2\epsilon\left[\varrho(\boldsymbol{\zeta}(\ell) - \mathsf{D} - \epsilon\bar{G}^2/2 - 2\epsilon DG\bar{G}\right]. \quad (64)$$

The square-bracketed quantity in (64) is positive due to the assumption of Case 1. Then (64) implies that $\|\boldsymbol{\zeta}(\ell+1) - \boldsymbol{\zeta}^*\|^2 < \|\boldsymbol{\zeta}(\ell) - \boldsymbol{\zeta}^*\|^2$, and the desired relation holds for $\ell+1$.

Case 2: $\varrho(\boldsymbol{\zeta}(\ell)) \leq \mathsf{D} + \epsilon G^2/2 + 2\epsilon DG\bar{G}$. It follows from (50), the nonexpansive property of the projection, the triangle inequality, and the bound (51) that

$$\|\boldsymbol{\zeta}(\ell+1) - \boldsymbol{\zeta}^*\| \leq \|\boldsymbol{\zeta}(\ell) + \epsilon\left(\check{\boldsymbol{g}}(t) + \check{\boldsymbol{v}}(\tau(\ell))\right) - \boldsymbol{\zeta}^*\| \quad (65a)$$
$$\leq \|\boldsymbol{\zeta}(\ell)\| + \|\boldsymbol{\zeta}^*\| + \epsilon G \quad (65b)$$

Next, a bound on $\|\boldsymbol{\zeta}(\ell)\|$ is developed. Specifically, it holds due to the definition of the dual function [cf. (38)] that

$$\varrho(\boldsymbol{\zeta}(\ell)) = \max_{\mathbf{y}\in\mathcal{B}_{\mathbf{y}}, \mathbf{p}\in\mathcal{P}} \left\{f(\mathbf{y}) - \boldsymbol{\zeta}^T(\ell)\big(\mathbf{g}(\mathbf{y}) + \mathbb{E}[\mathbf{v}(\mathbf{p}(\mathbf{h}), \mathbf{h})]\big)\right\}$$
$$\geq f(\mathbf{y}') - \boldsymbol{\zeta}^T(\ell)\big(\mathbf{g}(\mathbf{y}') + \mathbb{E}[\mathbf{v}(\mathbf{p}'(\mathbf{h}), \mathbf{h})]\big). \quad (66)$$

Rewriting the inner product in (66) using the entries of the corresponding vectors and substituting (62) into (66) using $\boldsymbol{\zeta} \geq \boldsymbol{0}$, it follows that

$$\delta \sum_{i=1}^{d} \boldsymbol{\zeta}_i(\ell) \leq -\sum_{i=1}^{d} \boldsymbol{\zeta}_i^T(\ell)\big(\mathbf{g}_i(\mathbf{y}') + \mathbb{E}[\mathbf{v}_i(\mathbf{p}'(\mathbf{h}), \mathbf{h})]\big)$$
$$\leq \varrho(\boldsymbol{\zeta}(\ell)) - f(\mathbf{y}'). \quad (67)$$

Using $\|\boldsymbol{\zeta}(\ell)\| \leq \sum_{i=1}^{d} \boldsymbol{\zeta}_i(\ell)$ into (67), the following bound is obtained:

$$\|\boldsymbol{\zeta}(\ell)\| \leq \frac{1}{\delta}(\varrho(\boldsymbol{\zeta}(\ell)) - f(\mathbf{y}')). \quad (68)$$

Introducing (68) into (65b) and using the assumption of Case 2, the desired relation (63) holds for $\ell+1$.

b) Set $\boldsymbol{\theta} = \boldsymbol{\zeta}^*$ and $\varrho(\boldsymbol{\theta}) = \varrho(\boldsymbol{\zeta}^*) = \mathsf{D}$ in (52c):

$$\|\boldsymbol{\zeta}(\ell+1) - \boldsymbol{\zeta}^*\|^2 \leq \|\boldsymbol{\zeta}(\ell) - \boldsymbol{\zeta}^*\|^2 + \epsilon^2 G^2 + 4\epsilon^2 DG\bar{G}$$
$$+ 2\epsilon\left[\mathsf{D} - \varrho(\boldsymbol{\zeta}(\ell))\right]. \quad (69)$$

Summing the latter for $\ell = 1, \ldots, s$, and introducing the quantity $\min_{1\leq\ell\leq s} \varrho(\boldsymbol{\zeta}(\ell))$, it follows that

$$\|\boldsymbol{\zeta}(\ell+1) - \boldsymbol{\zeta}^*\|^2 \leq \|\boldsymbol{\zeta}(1) - \boldsymbol{\zeta}^*\|^2 + s\epsilon^2 G^2 + 4s\epsilon^2 DG\bar{G}$$
$$+ 2s\epsilon\mathsf{D} - 2s\epsilon \min_{1\leq\ell\leq s} \varrho(\boldsymbol{\zeta}(\ell)). \quad (70)$$

Substituting the left-hand side of (70) with 0, rearranging the resulting inequality, and dividing by $2\epsilon s$, we obtain

$$\min_{1\leq\ell\leq s} \varrho(\boldsymbol{\zeta}(\ell)) \leq \mathsf{D} + \frac{\epsilon G^2}{2} + 2\epsilon DG\bar{G} + \frac{\|\boldsymbol{\zeta}(1) - \boldsymbol{\zeta}^*\|^2}{2\epsilon s}. \quad (71)$$

Now, note that $\lim_{s\to\infty} \min_{1\leq\ell\leq s} \varrho(\boldsymbol{\zeta}(\ell))$ exists, because $\min_{1\leq\ell\leq s} \varrho(\boldsymbol{\zeta}(\ell))$ is monotone decreasing in s and lower-bounded by D, which is finite. Moreover, $\lim_{s\to\infty} \|\boldsymbol{\zeta}(1) - \boldsymbol{\zeta}^*\|^2/(2\epsilon s) = 0$, because $\boldsymbol{\zeta}^*$ is bounded. Thus, taking the limit as $s \to \infty$ in (71), yields (32). $\qquad\square$

Note that the sequence of Lagrange multipliers in the synchronous algorithm (48c) is bounded. This was shown for convex primal problems in [25, Lemma 3]. Interestingly, the proof also applies in the present case since AS1-AS4 hold and

imply finite optimal P = D. Furthermore, Proposition 2 for the synchronous method follows from [27, Prop. 8.2.3], [22].

Next, the convergence of primal variables through running averages is considered. The following lemma collects the intermediate results for the averaged sequence $\{\bar{\mathbf{y}}(s)\}$ [cf. (24)], and is used to establish convergence for the generic problem (36) with asynchronous subgradient updates as in (50). Note that $\bar{\mathbf{y}}(s) \in \mathcal{B}_{\mathbf{y}}$, $s \geq 1$, because (24) represents a convex combination of the points $\{\mathbf{y}(1), \ldots, \mathbf{y}(s)\}$.

Lemma 3. *Under AS1-AS5 with $\boldsymbol{\zeta}^*$ denoting an optimal Lagrange multiplier vector, there exists a sequence $\{\mathring{\mathbf{p}}(.; s)\}$ in $\mathcal{P}$ such that for any $s \in \mathbb{N}$, it holds that*

a) $\left\| \left[\mathbf{g}(\bar{\mathbf{y}}(s)) + \mathbb{E}\left[\mathbf{v}(\mathring{\mathbf{p}}(\mathbf{h}; s), \mathbf{h}) \right] \right]^+ \right\| \leq \dfrac{\|\boldsymbol{\zeta}(s+1)\|}{\epsilon s}$ (72a)

b) $f(\bar{\mathbf{y}}(s)) \geq \mathsf{D} - \dfrac{\|\boldsymbol{\zeta}(1)\|^2}{2\epsilon s} - \dfrac{\epsilon G^2}{2} - 2\epsilon D G \bar{G}$ (72b)

c) $f(\bar{\mathbf{y}}(s)) \leq \mathsf{D} + \|\boldsymbol{\zeta}^*\| \left\| \left[\mathbf{g}(\bar{\mathbf{y}}(s)) + \mathbb{E}[\mathbf{v}(\mathring{\mathbf{p}}(\mathbf{h}; s), \mathbf{h})] \right]^+ \right\|$. (72c)

Eq. (72a) is an upper bound on the constraint violation, while (72b) and (72c) provide lower and upper bounds on the objective function at $\bar{\mathbf{y}}(s)$. Lemma 3 relies on Lemma 1 and the fact that the averaged sequence $\{\bar{\mathbf{y}}(s)\}$ is generated from maximizers of the Lagrangian $\{\mathbf{y}(\ell)\}$ that are *not* outdated.

Proof of Lemma 3: a) It follows from (50) that

$$\boldsymbol{\zeta}(\ell+1) \geq \boldsymbol{\zeta}(\ell) + \epsilon \left(\mathring{\boldsymbol{g}}(\ell) + \check{\boldsymbol{v}}(\tau(\ell)) \right).$$ (73)

Summing (73) over $\ell = 1, \ldots, s$, using $\boldsymbol{\zeta}(1) \geq \mathbf{0}$, and dividing by $2\epsilon s$, it follows that

$$\frac{1}{s} \sum_{\ell=1}^{s} \mathring{\boldsymbol{g}}(\ell) + \frac{1}{s} \sum_{\ell=1}^{s} \check{\boldsymbol{v}}(\tau(\ell)) \leq \frac{\boldsymbol{\zeta}(s+1)}{\epsilon s}.$$ (74)

Now, recall the definitions of the subgradients $\mathring{\boldsymbol{g}}(\ell)$ and $\check{\boldsymbol{v}}(\tau(\ell))$ in (49). Due to the convexity of $\mathbf{g}(\cdot)$, it holds that

$$\mathbf{g}(\bar{\mathbf{y}}(s)) \leq \frac{1}{s} \sum_{\ell=1}^{s} \mathbf{g}(\mathbf{y}(\ell)) = \frac{1}{s} \sum_{\ell=1}^{s} \mathring{\boldsymbol{g}}(\ell).$$ (75)

Due to Lemma 1, there exists $\mathring{\mathbf{p}}(\mathbf{h}; s)$ in $\mathcal{P}$ such that

$$\mathbb{E}[\mathbf{v}(\mathring{\mathbf{p}}(\mathbf{h}; s), \mathbf{h})] = \frac{1}{s} \sum_{\ell=1}^{s} \mathbb{E}[\mathbf{v}(\mathbf{p}(\mathbf{h}; \tau(\ell)), \mathbf{h})] = \frac{1}{s} \sum_{\ell=1}^{s} \check{\boldsymbol{v}}(\tau(\ell)).$$ (76)

Combining (74), (75), and (76), it follows that

$$\mathbf{g}(\bar{\mathbf{y}}(s)) + \mathbb{E}[\mathbf{v}(\mathring{\mathbf{p}}(\mathbf{h}; s), \mathbf{h})] \leq \frac{\boldsymbol{\zeta}(s+1)}{\epsilon s}.$$ (77)

Using $\boldsymbol{\zeta}(s+1) \geq \mathbf{0}$ and the fact that $[.]^+$ is a nonnegative vector, (72a) follows easily from (77).

b) Due to the concavity of $f(\cdot)$, it holds that $f(\bar{\mathbf{y}}(s)) \geq \frac{1}{s} \sum_{\ell=1}^{s} f(\mathbf{y}(\ell))$. Adding and subtracting the same terms, $\boldsymbol{\zeta}^T(\ell) \mathbf{g}(\mathbf{y}(\ell))$ and $\boldsymbol{\zeta}^T(\ell) \mathbb{E}[\mathbf{v}(\mathbf{p}(\mathbf{h}; \tau(\ell)), \mathbf{h})]$ for $\ell = 1, \ldots, s$, to the right-hand side of the latter, and using $f(\mathbf{y}(\ell)) - \boldsymbol{\zeta}^T(\ell) \mathbf{g}(\mathbf{y}(\ell)) = \psi(\boldsymbol{\zeta}(\ell))$ [cf. (48a) and (38)], it follows that

$$f(\bar{\mathbf{y}}(s)) \geq \frac{1}{s} \sum_{\ell=1}^{s} \left[\psi(\boldsymbol{\zeta}(\ell)) - \boldsymbol{\zeta}^T(\ell) \mathbb{E}[\mathbf{v}(\mathbf{p}(\mathbf{h}; \tau(\ell)), \mathbf{h})] \right]$$

$$+ \frac{1}{s} \sum_{\ell=1}^{s} \boldsymbol{\zeta}^T(\ell) \left(\mathbf{g}(\mathbf{y}(\ell)) + \mathbb{E}[\mathbf{v}(\mathbf{p}(\mathbf{h}; \tau(\ell)), \mathbf{h})] \right).$$ (78)

Now recall that $\mathbb{E}[\mathbf{v}(\mathbf{p}(\mathbf{h}; \tau(\ell)), \mathbf{h})] = \check{\mathbf{v}}(\tau(\ell))$ [cf. (49b)]. Thus, it holds that

$$- \boldsymbol{\zeta}^T(\ell) \mathbb{E}[\mathbf{v}(\mathbf{p}(\mathbf{h}; \tau(\ell)), \mathbf{h})]$$
$$= - \boldsymbol{\zeta}^T(\tau(\ell)) \check{\mathbf{v}}(\tau(\ell)) + \check{\mathbf{v}}^T(\tau(\ell)) [\boldsymbol{\zeta}(\tau(\ell)) - \boldsymbol{\zeta}(\ell)].$$ (79)

The first term in the right-hand side of (79) is $\phi(\boldsymbol{\zeta}(\tau(\ell)))$ ([cf. (48b) and (38)]. The second term can be lower-bounded using Lemma 2(a) with $\boldsymbol{\theta} = \boldsymbol{\zeta}(\tau(\ell))$. Then, (79) becomes

$$- \boldsymbol{\zeta}^T(\ell) \mathbb{E}[\mathbf{v}(\mathbf{p}(\mathbf{h}; \tau(\ell)), \mathbf{h})] \geq \phi(\boldsymbol{\zeta}(\ell)) - 2\epsilon D G \bar{G}$$ (80)

Using (80) into (78) and $\psi(\boldsymbol{\zeta}(\ell)) + \phi(\boldsymbol{\zeta}(\ell)) = \varrho(\boldsymbol{\zeta}(\ell)) \geq \mathsf{D}$, it follows that

$$f(\bar{\mathbf{y}}(s)) \geq \mathsf{D} - 2\epsilon D G \bar{G}$$

$$+ \frac{1}{s} \sum_{\ell=1}^{s} \boldsymbol{\zeta}^T(\ell) \left(\mathbf{g}(\mathbf{y}(\ell)) + \mathbb{E}[\mathbf{v}(\mathbf{p}(\mathbf{h}; \tau(\ell)), \mathbf{h})] \right).$$ (81)

Moreover, it follows from (50) and the nonexpansive property of the projection that

$$\|\boldsymbol{\zeta}(\ell+1)\|^2 \leq \|\boldsymbol{\zeta}(\ell)\|^2$$
$$+ 2\epsilon \boldsymbol{\zeta}^T(\ell) \left(\mathbf{g}(\mathbf{y}(\ell)) + \mathbb{E}[\mathbf{v}(\mathbf{p}(\mathbf{h}; \tau(\ell)), \mathbf{h})] \right)$$
$$+ \epsilon^2 \|\mathbf{g}(\mathbf{y}(\ell)) + \mathbb{E}[\mathbf{v}(\mathbf{p}(\mathbf{h}; \ell), \mathbf{h})]\|^2.$$ (82)

Summing (82) for $\ell = 1, \ldots, s$, dividing by $2\epsilon s$, and introducing the bound (51) on the subgradient norm yield

$$\frac{1}{s} \sum_{\ell=1}^{s} \boldsymbol{\zeta}^T(\ell) \left(\mathbf{g}(\mathbf{y}(\ell)) + \mathbb{E}[\mathbf{v}(\mathbf{p}(\mathbf{h}; \ell), \mathbf{h})] \right)$$

$$\geq - \frac{\epsilon G^2}{2} + \frac{\|\boldsymbol{\zeta}(s+1)\|^2 - \|\boldsymbol{\zeta}(1)\|^2}{2\epsilon s}.$$ (83)

Using (83) into (81) together with $\|\boldsymbol{\zeta}(s+1)\|^2 \geq 0$, one arrives readily at (72b).

c) Let $\boldsymbol{\zeta}^*$ be an optimal dual solution. It holds that

$$f(\bar{\mathbf{y}}(s)) = f(\bar{\mathbf{y}}(s)) - \boldsymbol{\zeta}^{*T} \left(\mathbf{g}(\bar{\mathbf{y}}(s)) + \mathbb{E}[\mathbf{v}(\mathring{\mathbf{p}}(\mathbf{h}; s), \mathbf{h})] \right)$$
$$+ \boldsymbol{\zeta}^{*T} \left(\mathbf{g}(\bar{\mathbf{y}}(s)) + \mathbb{E}[\mathbf{v}(\mathring{\mathbf{p}}(\mathbf{h}; s), \mathbf{h})] \right)$$ (84)

where $\mathring{\mathbf{p}}(\mathbf{h}; s)$ was defined in part (a) [cf. (76)].

By the definitions of D and $\boldsymbol{\zeta}^*$ [cf. (40)], and the dual function [cf. (38)], it holds that

$$\mathsf{D} = \varrho(\boldsymbol{\zeta}^*) = \max_{\mathbf{y} \in \mathcal{B}_{\mathbf{y}}, \mathbf{p} \in \mathcal{P}} \mathcal{L}(\mathbf{y}, \mathbf{p}, \boldsymbol{\zeta}^*) \geq \mathcal{L}(\bar{\mathbf{y}}, \mathring{\mathbf{p}}, \boldsymbol{\zeta}^*).$$ (85)

Substituting the latter into (84), it follows that

$$f(\bar{\mathbf{y}}(s)) \leq \mathsf{D} + \boldsymbol{\zeta}^{*T} \left(\mathbf{g}(\bar{\mathbf{y}}(s)) + \mathbb{E}[\mathbf{v}(\mathring{\mathbf{p}}(\mathbf{h}; s), \mathbf{h})] \right).$$ (86)

Because $\boldsymbol{\zeta}^* \geq \mathbf{0}$ and $\boldsymbol{\theta} \leq [\boldsymbol{\theta}]^+$ for all $\boldsymbol{\theta}$, (86) implies that

$$f(\bar{\mathbf{y}}(s)) \leq \mathsf{D} + \boldsymbol{\zeta}^{*T} \left[\mathbf{g}(\bar{\mathbf{y}}(s)) + \mathbb{E}[\mathbf{v}(\mathring{\mathbf{p}}(\mathbf{h}; s), \mathbf{h})] \right]^+.$$ (87)

Applying the Cauchy-Schwartz inequality to the latter, (72c) follows readily. $\square$

Using Lemma 3, the main convergence results for the synchronous and asynchronous subgradient methods are given correspondingly by Propositions 3 and 5, after substituting

$$\mathbf{q}(\bar{\mathbf{y}}(s), \mathring{\mathbf{p}}(\mathbf{h}; s)) = \mathbf{g}(\bar{\mathbf{y}}(s)) + \mathbb{E}[\mathbf{v}(\mathring{\mathbf{p}}(\mathbf{h}; s))].$$ (88)

Proof of Proposition 5: a) Take limits on both sides of (72a) as $s \to \infty$, and use the boundedness of $\{\zeta(s)\}$.

b) Using $\mathsf{P} = \mathsf{D}$ and taking the $\liminf$ in (72b), we obtain (34a). Moreover, using $\mathsf{P} = \mathsf{D}$, (72a), the boundedness of $\|\zeta^*\|$, and taking $\limsup$ in (72c), (34b) follows. ∎

REFERENCES

[1] R. W. Yeung, S.-Y. R. Li, N. Cai, and Z. Zhang, "Network coding theory," *Foundations and Trends in Communications and Information Theory*, vol. 2, no. 4 and 5, pp. 241–381, 2005.

[2] P. Chou and Y. Wu, "Network coding for the Internet and wireless networks," *IEEE Signal Process. Mag.*, vol. 24, no. 5, pp. 77–85, Sep. 2007.

[3] K. Bharath-Kumar and J. Jaffe, "Routing to multiple destinations in computer networks," *IEEE Trans. Commun.*, vol. COM-31, no. 3, pp. 343–351, Mar. 1983.

[4] R. Ahlswede, N. Cai, S.-Y. R. Li, and R. Yeung, "Network information flow," *IEEE Trans. Inf. Theory*, vol. 46, no. 4, pp. 1204–1216, July 2000.

[5] D. S. Lun, N. Ratnakar, M. Medard, R. Koetter, D. R. Karger, T. Ho, E. Ahmed, and F. Zhao, "Minimum-cost multicast over coded packet networks," *IEEE Trans. Inf. Theory*, vol. 52, no. 6, pp. 2608–2623, June 2006.

[6] X. Lin, N. B. Shroff, and R. Srikant, "A tutorial on cross-layer optimization in wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 8, pp. 1452–1463, Aug. 2006.

[7] L. Georgiadis, M. J. Neely, and L. Tassiulas, "Resource allocation and cross-layer control in wireless networks," *Foundations and Trends in Networking*, vol. 1, no. 1, pp. 1–144, 2006.

[8] M. Chiang, S. H. Low, A. R. Calderbank, and J. C. Doyle, "Layering as optimization decomposition: A mathematical theory of network architectures," *Proc. IEEE*, vol. 95, no. 1, pp. 255–312, Jan. 2007.

[9] D. S. Lun, M. Medard, R. Koetter, and M. Effros, "On coding for reliable communication over packet networks," in *Proc. 42nd Annu. Allerton Conf. Communication, Control, and Computing*, Monticello, IL, Sep. 2004, pp. 20–29.

[10] P. A. Chou, Y. Wu, and K. Jain, "Practical network coding," in *Proc. 41st Annu. Allerton Conf. Communication, Control, and Computing*, Monticello, IL, Oct. 2003, pp. 40–49.

[11] Y. Xi and E. Yeh, "Distributed algorithms for minimum cost multicast with network coding," *IEEE/ACM Trans. Netw.*, to be published.

[12] T. Cui, L. Chen, and T. Ho, "Energy efficient opportunistic network coding for wireless networks," in *Proc. IEEE INFOCOM*, Phoenix, AZ, Apr. 2008, pp. 1022–1030.

[13] D. Traskov, M. Heindlmaier, M. Medard, R. Koetter, and D. S. Lun, "Scheduling for network coded multicast: A conflict graph formulation," in *IEEE GLOBECOM Workshops*, New Orleans, LA, Nov.–Dec. 2008.

[14] M. H. Amerimehr, B. H. Khalaj, and P. M. Crespo, "A distributed cross-layer optimization method for multicast in interference-limited multihop wireless networks," *EURASIP J. Wireless Commun. and Networking*, vol. 2008, June 2008.

[15] Y. Wu and S.-Y. Kung, "Distributed utility maximization for network coding based multicasting: A shortest path approach," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 8, pp. 1475–1488, Aug. 2006.

[16] Y. Wu, M. Chiang, and S.-Y. Kung, "Distributed utility maximization for network coding based multicasting: A critical cut approach," in *Proc. 4th Int. Symp. Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks*, Boston, MA, Apr. 2006.

[17] Z. Li and B. Li, "Efficient and distributed computation of maximum multicast rates," in *Proc. IEEE INFOCOM*, Miami, FL, Mar. 2005, pp. 1618–1628.

[18] L. Chen, T. Ho, S. H. Low, M. Chiang, and J. C. Doyle, "Optimization based rate control for multicast with network coding," in *Proc. IEEE INFOCOM*, Anchorage, AK, May 2007, pp. 1163–1171.

[19] D. Li, X. Lin, W. Xu, Z. He, and J. Lin, "Rate control for network coding based multicast: a hierarchical decomposition approach," in *Proc. 5th Int. Conf. Wireless Communications and Mobile Computing*, June 2009, pp. 181–185.

[20] X. Yan, M. J. Neely, and Z. Zhang, "Multicasting in time-varying wireless networks: Cross-layer dynamic resource allocation," in *IEEE Int. Symp. Information Theory*, Nice, France, June 2007, pp. 2721–2725.

[21] T. Ho and H. Viswanathan, "Dynamic algorithms for multicast with intra-session network coding," *IEEE Trans. Inf. Theory*, vol. 55, no. 2, pp. 797–815, Feb. 2009.

[22] A. Ribeiro and G. B. Giannakis, "Separation theorems of wireless networking," *IEEE Trans. Inf. Theory*, vol. 56, no. 9, Sep. 2010.

[23] N. Gatsis, A. Ribeiro, and G. Giannakis, "A class of convergent algorithms for resource allocation in wireless fading networks," *IEEE Trans. Wireless Commun.*, vol. 9, no. 5, pp. 1808–1823, may 2010.

[24] K. C. Kiwiel and P. O. Lindberg, "Parallel subgradient methods for convex optimization," in *Inherently Parallel Algorithms in Feasibility and Optimization*, D. Butnariu, Y. Censor, and S. Reich, Eds. Amsterdam, Netherlands: Elsevier Science B.V., 2001, pp. 335–344.

[25] A. Nedić and A. Ozdaglar, "Approximate primal solutions and rate analysis for dual subgradient methods," *SIAM J. Optim.*, vol. 19, no. 4, pp. 1757–1780, 2009.

[26] Z.-Q. Luo and S. Zhang, "Dynamic spectrum management: Complexity and duality," *IEEE J. Sel. Topics Signal Process.*, vol. 2, no. 1, pp. 57–73, Feb. 2008.

[27] D. P. Bertsekas, A. Nedić, and A. Ozdaglar, *Convex Analysis and Optimization*. Belmont, MA: Athena Scientific, 2003.

[28] M. Heindlmaier, D. Traskov, R. Kotter, and M. Medard, "Scheduling for network coded multicast: A distributed approach," in *IEEE GLOBECOM Workshops*, Honolulu, HI, Nov. 2009, pp. 1–6.

[29] D. P. Bertsekas, *Nonlinear Programming*. Belmont, MA: Athena Scientific, 1999.

[30] D. Blackwell, "On a theorem of Lyapunov," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 112–114, Mar. 1951.

[31] N. Dinculeanu, *Vector Measures*. Oxford, U.K.: Pergamon Press, 1967.

[32] A. Nedić and D. P. Bertsekas, "The effect of deterministic noise in subgradient methods," *Math. Programming*, 2009. [Online]. Available: http://dx.doi.org/10.1007/s10107-008-0262-5

PLACE PHOTO HERE

Ketan Rajawat (S'06) received his B.Tech.and M.Tech.degrees in Electrical Engineering from Indian Institute of Technology Kanpur in 2007. Since August 2007, he has been working towards his Ph.D. degree at the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis. His current research focuses on network coding and optimization in wireless networks.

PLACE PHOTO HERE

Nikolaos Gatsis received the Diploma degree in electrical and computer engineering from the University of Patras, Patras, Greece in 2005 with honors. Since September 2005, he has been working toward the Ph.D. degree with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN. His research interests include cross-layer designs, resource allocation, and signal processing for wireless networks.

PLACE PHOTO HERE

G. B. Giannakis (F'97) received his Diploma in Electrical Engr. from the Ntl. Tech. Univ. of Athens, Greece, 1981. From 1982 to 1986 he was with the Univ. of Southern California (USC), where he received his MSc. in Electrical Engineering, 1983, MSc. in Mathematics, 1986, and Ph.D. in Electrical Engr., 1986. Since 1999 he has been a professor with the Univ. of Minnesota, where he now holds an ADC Chair in Wireless Telecommunications in the ECE Department, and serves as director of the Digital Technology Center.

His general interests span the areas of communications, networking and statistical signal processing - subjects on which he has published more than 300 journal papers, 500 conference papers, two edited books and two research monographs. Current research focuses on compressive sensing, cognitive radios, network coding, cross-layer designs, mobile ad hoc networks, wireless sensor and social networks. He is the (co-) inventor of 18 to 20 patents issued, and the (co-) recipient of seven paper awards from the IEEE Signal Processing (SP) and Communications Societies, including the G. Marconi Prize Paper Award in Wireless Communications. He also received Technical Achievement Awards from the SP Society (2000), from EURASIP (2005), a Young Faculty Teaching Award, and the G. W. Taylor Award for Distinguished Research from the University of Minnesota. He is a Fellow of EURASIP, and has served the IEEE in a number of posts, including that of a Distinguished Lecturer for the IEEE-SP Society.